Histoire évolutive et propagation de la tuberculose à échelle planétaire : vers une approche intégrée combinant la génomique des populations et le typage multi-locus

Maxime Barbier

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'École Pratique des Hautes Études

Histoire évolutive et propagation de la tuberculose à échelle planétaire : vers une approche intégrée combinant la génomique des populations et le typage multi-locus

École doctorale de l'EPHE – ED 472

Spécialité : BIODIVERSITÉ, GÉNÉTIQUE ET ÉVOLUTION

Soutenue par :
Maxime BARBIER
le 11 décembre 2017

Dirigée par :
Thierry WIRTH

COMPOSITION DU JURY :

M. Roland BROSCH
Institut Pasteur de Paris
Président du jury - Rapporteur

M. Sylvain BRISSE
Institut Pasteur de Paris
Rapporteur

M. Sébastien GAGNEUX
Swiss Tropical and Public Health Institute
Examinateur

M. Guillaume ACHAZ
Université Pierre et Marie Curie
Examinateur

M. Thierry WIRTH
École Pratique des Hautes Études
Directeur de thèse

THESE

Pour obtenir le grade de Docteur de l'Ecole Pratique
des Hautes Etudes

ED 472

Présentée et soutenue publiquement par

Maxime Barbier

Le 11 décembre 2017

Histoire évolutive et propagation de la tuberculose à échelle planétaire : vers une approche intégrée combinant la génomique des populations et le typage multi-locus

Membres du Jury :

Rapporteur 1 : **M. Roland Brosch**, directeur de recherche de l'Institut Pasteur de Paris

Rapporteur 2 : **M. Sylvain Brisse**, directeur de recherche de l'Institut Pasteur de Paris

Examinateur 1 : **M. Sébastien Gagneux**, professeur au Swiss Tropical and Public Health Institute

Examinateur 2 : **M. Guillaume Achaz**, maître de conférence de l'Université Pierre et Marie Curie

Directeur de Thèse : **M. Thierry Wirth**, directeur d'étude de l'EPHE

Abstract

According to a 2015 WHO report, tuberculosis remains one of the top 10 causes of death worldwide. Despite considerable efforts by the United Nations to eradicate the disease by 2030, a global TB epidemic still persists. Its causative agent, the bacterium *Mycobacterium tuberculosis*, an obligate pathogen, has been plaguing humanity since it originated, and has coevolved with its main host, *Homo sapiens*, over thousands of years. Contemporary tuberculosis strains exhibit a structured phylogeographic pattern, carrying the genetic print of their geographic origin. The Koch bacillus infects and kills in large numbers, in poor and developing countries, where fragile health care systems, combined with high HIV prevalence, facilitate epidemic spread. In western countries, the major current threats are the multiplication and propagation of antibiotic resistant strains (MDR/XDR) coming predominantly from former Soviet republics.

In this thesis, I unravel the evolutionary history, propagation, and acquisition of drug resistance-conferring mutations in different settings, by implementing multiple genetic and genomic data sets. First, focusing on Central Asia, using whole genome sequencing and Bayesian statistics, I assess the effects of a treatment campaign on the development of MDR strains and highlight key mutations in successful strains. More importantly, the success of DOTs campaigns was compromised by the genetic make-up of these outbreak clades (pre-treatment low frequency resistance SNPs). Special attention was also given to a particular outbreak of MDR strains, i.e. the Russian W148 clone. I present its westward spatial and temporal propagation at a continental scale during the last century, and underline the key contribution of compensatory mutations in its epidemic success. However, tuberculosis does not only infect humans, but also has experienced successive mammalian host jumps. To decipher the adaptive constraints accompanying such secondary events, a systemic gene screen with selection signature-detecting algorithms was implemented to identify putative targets during diversifying selection.

Finally, novel mathematical tools and indices that reflect the epidemicity of a strain were developed, jumping from a population-driven approach to a strain specific one, with broader epidemiological applications. This allows us to correlate strain fitness with patient, lineage, and socio-economic information.

Résumé

D'après un rapport de l'OMS, la tuberculose reste en 2015 l'une des 10 premières causes de décès à l'échelle mondiale. De ce fait, en matière de santé, éradiquer la maladie à l'horizon 2030 est un des objectifs majeurs fixés par les Nations Unies. La bactérie responsable de cette infection, *Mycobacterium tuberculosis*, est un pathogène obligatoire dont l'origine et l'évolution sont intrinsèquement liées à celles de son hôte principal, *Homo sapiens*. En effet, les souches actuelles de tuberculose présentent, tout comme l'homme, une forte structure phylogénétique, trace de leur origine géographique. Les pays pauvres et en développement sont les plus touchés par l'épidémie globale, favorisée par des systèmes de santé défaillants et une haute prévalence du VIH. Les pays occidentaux ne sont pas épargnés, menacés par l'émergence de souches de plus en plus résistantes aux antibiotiques provenant en grande partie de l'ex URSS.

Au cours de cette thèse, j'analyse l'histoire évolutive, la propagation et l'acquisition de résistances aux antibiotiques de plusieurs épidémies de tuberculose en me basant sur des données génétiques et génomiques. Dans un premier temps je m'intéresse aux effets d'une campagne nationale de traitements en Asie Centrale sur le développement de souches multi-résistantes et met également en lumière le rôle clef de certaines mutations dans le succès des clones présentés. Ainsi cette campagne a été partiellement mise en échec par la présence de souches pré-résistantes, grâce à la survenue de mutations avant même la mise en place des traitements antibiotiques. Par la suite je me suis focalisé sur un clade particulier de souches multi-résistantes, le clone Russe W148. Je présente sa dispersion géographique et temporelle à travers l'Eurasie et démontre l'importance des mutations compensatoires dans son succès épidémique. De plus, la tuberculose ne touche pas seulement les hommes mais infecte également plusieurs autres mammifères. Afin d'appréhender les contraintes adaptatives accompagnants ces changements d'hôtes, j'ai effectué divers tests de sélection dans le but d'identifier les gènes impliqués.

Pour finir, nous avons développé un indice souche spécifique, permettant de mesurer le succès épidémique de celles-ci à un niveau individuel. Dans le cadre d'études épidémiologiques, cette mesure peut être croisée avec des informations sur le patient, la souche ou même socio-économiques.

Remerciements

Je souhaite tout d'abord remercier les membres de mon jury de thèse pour avoir accepté d'évaluer le manuscrit, Roland Brosch et Sylvain Brisse qui seront les rapporteurs,ainsi que Sébastien Gagneux et Guillaume Achaz, mes examinateurs.

Je souhaite également remercier toutes les personnes avec qui j'ai pu collaborer de près ou de loin durant ces 3 ans. Tout d'abord Thierry, mon directeur de thèse qui fut un très bon encadrant mais également un ami. Jean Philippe avec qui ça a été un plaisir de travailler durant une année, et tous nos collaborateurs sur la scène nationale et internationale, Philip Supply, Stefan Niemann et Matthias Merker. Je tiens à remercier également tous les membres passés ou encore présents au sein de l'équipe, Stefano, Arnaud, Pascaline et Claudie qui furent de chaleureux collègues, toujours à l'écoute et qui ont su me stimuler durant ces 3 dernières années.

Je remercie particulièrement Leslie pour sa présence et son soutien, ainsi que toute ma famille, mes parents, et les amis que j'ai pu côtoyer tout ce temps à Paris.

Enfin je tiens à remercier toutes les personnes qui assisteront à la soutenance de thèse et qui viendront au pot !

Sommaire

Introduction

Mycobacterium tuberculosis et pathogénèse de la tuberculose

L'agent étiologique de la Tuberculose est la bactérie *Mycobacterium tuberculosis.* Il s'agit d'une mycobactérie à croissance lente, elle a un temps de doublement compris entre 12 et 24h ; en culture sur milieu solide, des colonies apparaissent après 2 à 4 semaines. C'est un bacille aérobie strict qui ne forme pas de spore et n'est pas mobile (Delogu et al. 2013). Sa principale caractéristique est sa paroi, particulière aux mycobactéries, et dans le cas de *M. tuberculosis* elle joue un rôle majeur dans sa pathogénicité. Elle est formée d'un peptidoglycane (ou muréine) arabino-galactane et d'acides lipoteichoïques à laquelle vient s'ajouter une double couche lipidique à acides mycoliques sur la face externe, conduisant à une paroi très hydrophobe lui conférant une résistance importante aux macrophages. De plus au-delà de cette paroi une capsule est présente (Hoffmann et al. 2008). La tuberculose se transmet essentiellement par aérosols, lorsqu'un individu infecté tousse, il va expectorer des micro-gouttelettes de salive et de mucus dans lesquelles des bactéries sont présentes. Les personnes proches vont inhaler ces particules infectieuses en respirant et peuvent dans certains cas développer une tuberculose pulmonaire à leur tour. Une manière plus marginale d'acquisition de la maladie, et ce depuis la pasteurisation du lait, est de boire du lait provenant de vaches infectées par *Mycobacterium bovis*, une souche infectant principalement les bovins mais transmissible à l'homme, provoquant une tuberculose intestinale. Lorsque le bacille va pénétrer l'alvéole pulmonaire, lors de l'infection d'un hôte, la réponse immunitaire de ce dernier va permettre d'encapsuler la bactérie sous forme de granulomes (Figure 1), qui vont former une barrière physique contenant l'infection et bloquant la dissémination de la bactérie (Gengenbacher and Kaufmann 2012). La phase latente peut durer plusieurs dizaine d'années, c'est un processus dynamique entre la bactérie et le système immunitaire de l'hôte. Dans certains cas celui-ci ne développera jamais la maladie, dans d'autres la tuberculose latente se réactivera.

De nombreux facteurs favorisent cette réactivation comme le VIH, la malnutrition, le tabagisme, la pollution et l'alcoolisme (Kasner et al. 2013). Une dissémination hématogène des bacilles peut avoir lieu et entrainer une tuberculose extra-pulmonaire par la suite. Celle-ci peut toucher de nombreux organes. Le type de tuberculose variera en fonction de l'organe infecté. Parmi ceux-ci, il y a par exemple, la tuberculose miliaire qui survient lors de la

dispersion massive des bactéries dans le sang après la désagrégation d'un foyer tuberculeux. Elle entraine souvent l'infection de la moelle osseuse et d'autres organes. Les symptômes sont de la fièvre, une très grande fatigue, des malaises. Si la moelle osseuse est touchée une réaction de type leucémique peut survenir.

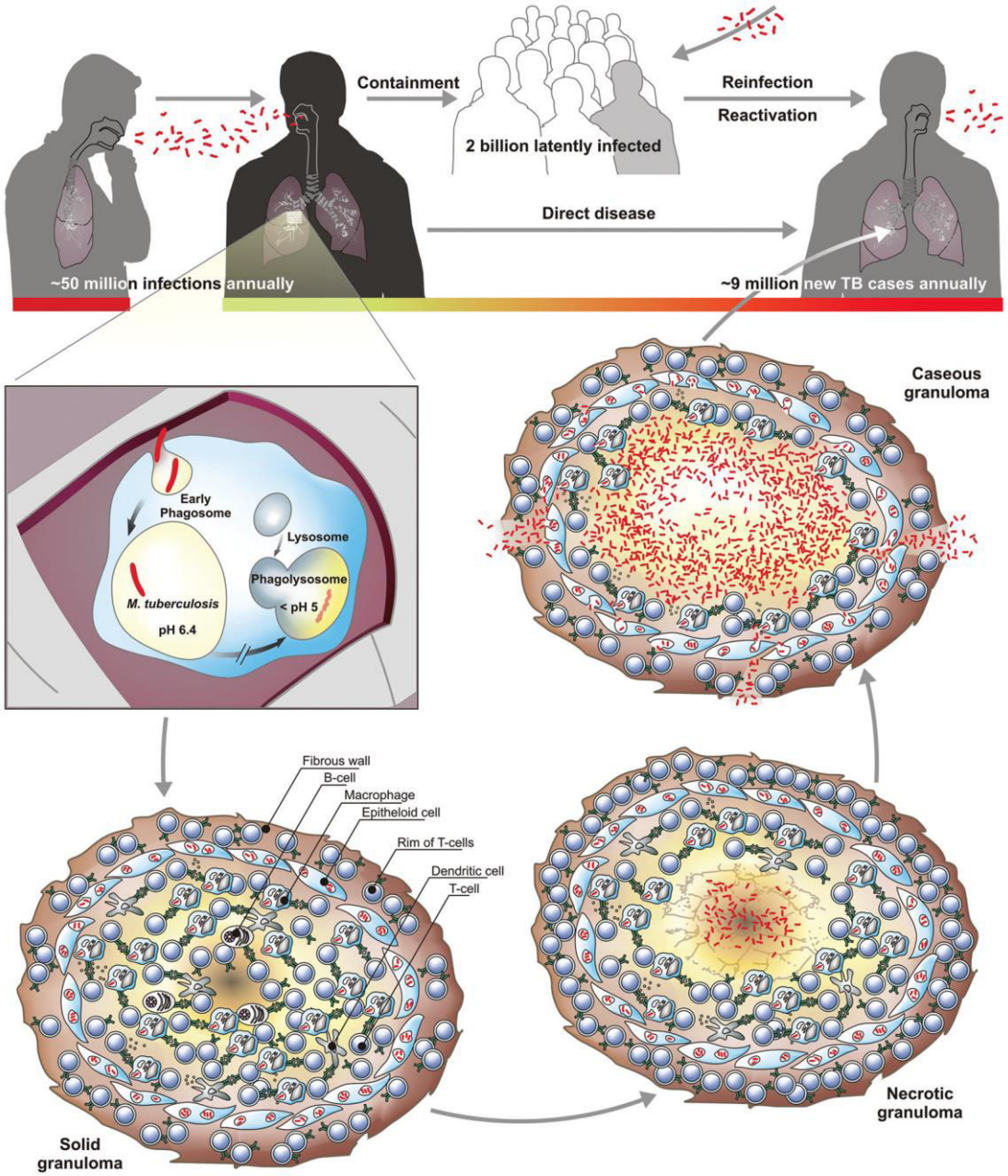


Figure 1 : Transmission et pathologie de la tuberculose tirée de Gengenbacher and Kaufmann 2012. La transmission de la tuberculose entre individus se fait par transmission aérienne de bacilles infectieux. Dans quelques cas seulement l'infection débouchera directement sur une tuberculose active. Via l'inhalation de gouttelettes, le pathogène va atteindre les voies respiratoires pulmonaires et être phagocyté par les macrophages alvéolaires. La cellule infectée va déclencher une réponse inflammatoire locale attirant des cellules mononucléaires et des lymphocytes T, entrainant la formation d'un granulome. C'est ce qu'on appelle le stade d'infection latente, qui sera maintenue chez les individus sains mais qui risqueront une réactivation au cours de leur vie. L'évolution du granulome (solide, nécrotique, caséeux) peut être plus ou moins rapide et atteint son climax au même moment que les différentes lésions apparaissant durant la tuberculose active. Le granulome caséeux perd sa solidité au cours de la désagrégation de son centre en une accumulation de débris cellulaire, le

caséum. *M. tuberculosis* se multiplie grandement, est libérée dans les voies respiratoires et est expectorée en aérosol contagieux.

La méningite tuberculeuse est la forme de tuberculose la plus grave, elle s'accompagne de la mortalité la plus élevée. Elle peut entrainer une céphalée persistante, un état de somnolence, des nausées et aller jusqu'à plonger le malade dans un coma. La lymphadénite tuberculeuse est quant à elle l'atteinte des ganglions cervicaux par la tuberculose. Elle entraîne un gonflement de ces ganglions jusqu'au stade où la peau peut s'ouvrir. La tuberculose ostéoarticulaire est une infection des articulations entrainant une arthrite chronique et des douleurs. On l'appelle maladie de Pott lorsque la colonne vertébrale est touchée, les vertèbres vont alors se tasser, entraînant une compression de la moelle épinière pouvant provoquer des problèmes neurologiques et une paraplégie. Il peut également se former un abcès sur la colonne vertébrale, laissant des traces identifiables sur les squelettes (Kumar et al. 2010).

MTBC (*Mycobacterium tuberculosis complex*)

On parle du MTBC pour désigner toute la diversité des souches proches de *M. tuberculosis,* provoquant la même maladie, ayant un phénotype et un génotype extrêmement proche (99.95% de similarité au niveau nucléotidique), mais pouvant être différenciées phylogénétiquement (Rodriguez-Campos et al. 2014). *M. tuberculosis* en fait donc partie et le MTBC peut être divisé en 7 lignées (Figure 2) (Comas et al. 2013). Parmi ces membres on compte *Mycobacterium africanum* qui infecte l'homme et qui est principalement confiné à Afrique de l'ouest (Winglee et al. 2016). Désormais *M. africanum* forme les lignées 5 et 6. En dehors de l'homme, la tuberculose touche de nombreuses espèces de mammifères, où la maladie prend des formes et symptômes voisins. Selon l'animal infecté un nom binomial différent est attribué à la mycobactérie responsable, dépendant de l'animal infecté. Cette nomenclature taxonomique a été secondairement validée par les données moléculaires. *M. bovis* est caractéristique des souches affectant les bovins (Karlson 1970), *Mycobacterium caprae* des souches affectant les caprins (Aranaz et al. 2003), *Mycobacterium microti* des souches affectant principalement les rongeurs (Boniotti et al. 2014). *Mycobacterium orygis* (van Ingen et al. 2012) des souches affectant les oryx, les gazelles et les antilopes. *Mycobacterium pinnipedii* quant à lui affecte les pinnipèdes (phoques, otaries) (Cousins et al. 2003). *Mycobacterium suricattae* et *Mycobacterium mungi* sont retrouvées respectivement chez les suricates et les mangoustes (Dippenaar et al. 2015; Alexander et al. 2010). Très récemment, une souche a également été retrouvée chez un chimpanzé (Coscolla et al. 2013), cette dernière semble phylogénétiquement proche de *M. africanum*.

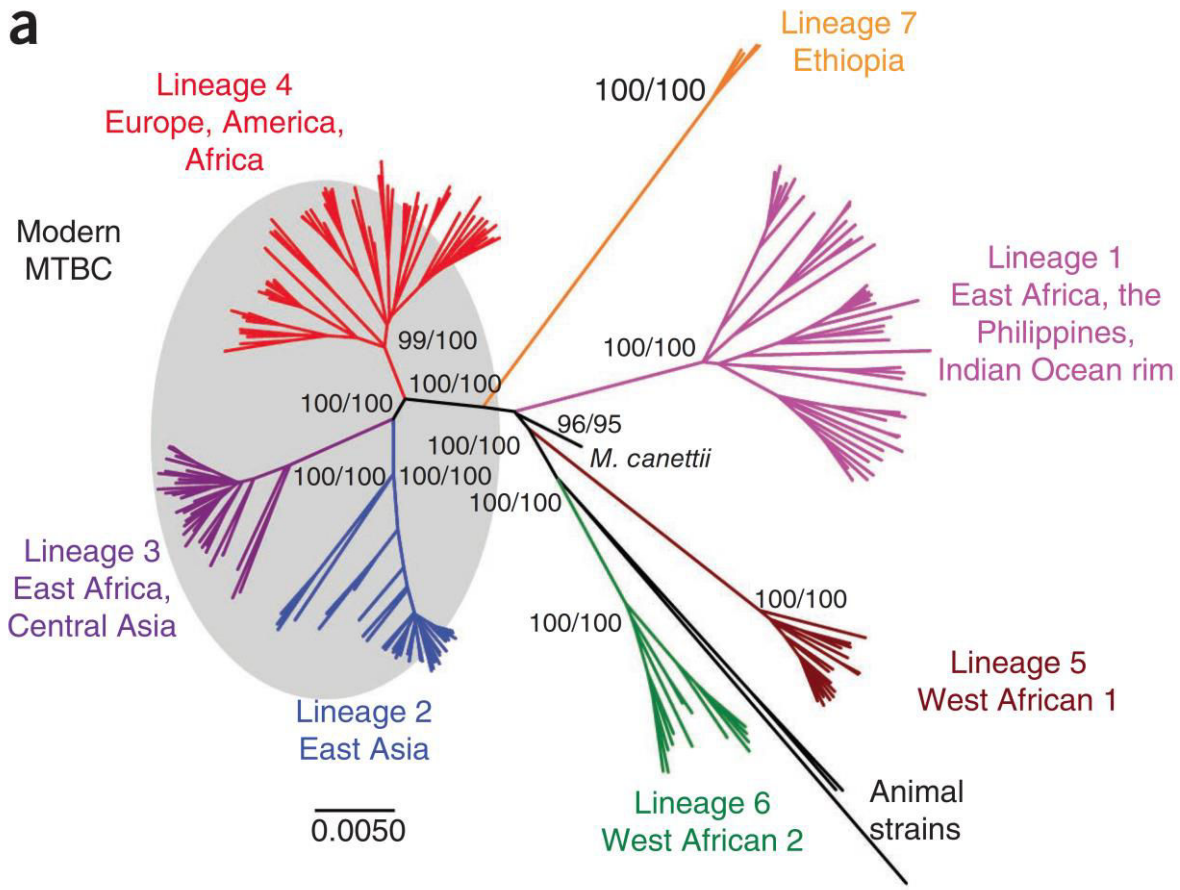


Figure 2 : Phylogénie en génome complet de 220 souches de tuberculose extrait de Comas et al. 2013. Les valeurs sur les branches principales indiquent leur support statistique, à gauche inféré par neighbor-joining et à droite par maximum likelihood. On peut voir les 7 lignées humaines ainsi que la lignée animale. *Mycobacterium canettii* représente l'ancêtre le plus vraisemblable du MTBC.

Homme et tuberculose au travers des âges

Le genre *Mycobacterium* aurait au moins 150 millions d'années et des traces de *M. tuberculosis* sont retrouvées dès le Paléolithique supérieur. La plupart sont observées sur des os dû aux lésions provoquées par la tuberculeuse ostéoarticulaire appelée également maladie de Pott (Pigrau-Serrallach and Rodriguez-Pardo 2013). On retrouve ces lésions particulières à la tuberculose et confirmées moléculairement sur des momies égyptiennes (Zink Albert et al. 2001), sur des squelettes datant de 9000 ans dans le croissant fertile (Hershkovitz et al. 2008) et même sur les restes d'un bison datant de 17000 ans en Amérique (Rothschild et al. 2001). Cependant l'apparition de *M. tuberculosis* est jugée plus ancienne encore et pourrait remonter à la domestication du feu par l'homme, qui réunissait les conditions idéales pour que son ancêtre, une souche bactérienne environnemental devienne un pathogène transmissible (Chisholm et al. 2016). Il n'y pas de preuve tangible en faveur de cette hypothèse mais il n'est

pas exclu que la tuberculose soit apparue avant l'Out-of-Africa, dès les premiers *Homo sapiens* comme détaillé dans le chapitre 1.

Les premiers témoignages historiques retrouvés sur la tuberculose datent du VIIème siècle avant JC, sur des tablettes assyriennes en argiles représentant des patients toussant du sang et vers 460 avant JC en Grèce antique, dans des textes d'Hippocrates où la maladie est nommée phtisie, signifiant consomption. Il s'agirait de l'infection la plus répandue (Smith 2003), survenant entre 18 et 35 ans et presque toujours mortelle. Etonnamment la phtisie était considérée comme résultant de la faiblesse mentale des personnes infectées et serait héréditaire, à l'exception notable d'Aristote qui l'identifia comme contagieuse. Dans l'empire romain, au V^{ème} siècle après JC, des remèdes tels que l'héliothérapie, des bains d'urine ou une alimentation riche en divers organes d'animaux étaient recommandés. Bien sûr ceux-ci étaient inefficaces. Au moyen-âge, les scrofules, ou écrouelles, inflammations des ganglions et tuméfactions au niveau du cou causées par une lymphadénite tuberculeuse, étaient appelées « mal royal » car on pensait que les rois de France et d'Angleterre pouvaient les guérir par le touché. Ce traitement, et d'autres tout aussi fantaisistes, demeuraient inefficaces, n'ayant aucun valeur thérapeutique réelle (Carlos et al. 2007).

Au XVII^{ème} siècle commença une épidémie de tuberculose sans précédent dans l'histoire en Europe et en Amérique du Nord. Nommée grande peste blanche, celle-ci dura plus de 200 ans et atteint son pic au début du XIX^{ème} siècle où on estime qu'un quart des Européens mourra de la tuberculose. Le développement des grandes villes à cette époque, induisant des densités de population jamais rencontrées auparavant, accompagné d'une insalubrité extrême, formaient les conditions idéales pour le développement de cette maladie épidémique aérotransmissible. La tuberculose devint la première cause de mortalité, celle-ci allant de 700 à 1000 pour 100000 personnes par années dans les grandes villes européennes et américaines (Daniel 2006). Suite aux grandes explorations coloniales de cette époque, la tuberculose se répandit à échelle planétaire, du moins en ce qui concerne les souches européennes. Elles causèrent ainsi des dégâts considérables au sein des populations indigènes des Amériques et d'Afrique. Dans certains camps de confinement des indigènes d'Amérique du Nord par exemple, le taux de mortalité a pu atteindre 9000 pour 100000 personnes (Bates and Stead 1993). L'amélioration de l'hygiène, des conditions de vie, ainsi que le développement du tout-à-l'égout dans les grandes villes durant la seconde partie du XIX^{ème} entrainèrent une baisse de la mortalité due à la tuberculose.

Dans les pays développés la déclive se poursuivit, aidée par les avancées médicales et une meilleure compréhension de la nature même de la tuberculose au cours de ces deux siècles. La découverte de « tubercules » dans les organes, bien souvent les poumons, des patients souffrant de consomption par Franciscus Sylvius de la Böe (1614-1672) et la preuve apportée par Gaspard Laurent Bayle (1774-1816) que ces derniers étaient la cause de la maladie ont permis un saut quantique dans la perception médicale de la peste blanche. C'est à partir de ce moment qu'on parla de tuberculose pour unifier deux maladies qu'on pensait jusqu'alors distinctes : La scrofule, lors de tuberculose extra-pulmonaire, et la phtisie pour la forme pulmonaire. Une étape cruciale dans la compréhension de l'épidémiologie de cette maladie fut de découvrir sa nature infectieuse. En effet, la tuberculose était majoritairement vue comme une maladie héréditaire en ces temps. Benjamin Marten (1704-1722) fut l'un des premiers à formuler l'hypothèse que la tuberculose était causée par des « minuscules créatures vivantes ». Cependant, Jean-Antoine Villemin (1827-1892) fut le premier à en démontrer la nature infectieuse *in situ* en infectant une vache puis un lapin grâce à du liquide purulent prélever dans la cavité tuberculeuse d'un patient décédé. Il en vint à la conclusion que la maladie était bien provoquée par des microorganismes et qu'elle était transmissible. La découverte majeure fût la confirmation que la tuberculose était bien causée par un microorganisme, un bacille plus précisément. En 1882, Robert Koch (1843-1910) fit part de sa découverte au cours de sa célèbre présentation « Die Aetiologie der Tuberculose » (Koch 1882). Il y présenta notamment un nouveau moyen de faire des cultures pures de bactéries mais aussi de les colorer grâce à du bleu de méthylène. Il montra à l'audience que dans chaque infection de tuberculose la bactérie était présente mais aussi qu'elle en était la cause en réinfectant des animaux sains. Grâce à sa technique de coloration utilisant du bleu de méthylène, il pouvait dévoiler le germe au milieu des tissus infectés, la bactérie se colorant après un traitement à la vésuvine (Kaufmann and Schaible 2005). Se basant sur ces travaux et les travaux de Friedrich Loeffler (1852-1915) sur la diphtérie, il énonça les postulats de *Koch-Henle* pour définir un pathogène microbien. Ce sont 4 critères indispensables pour désigner un microbe comme cause d'une maladie :

i) *L'organisme doit être trouvé dans tous les animaux souffrant de la maladie mais pas dans les animaux sains.*

ii) L'organisme doit être isolé à partir d'un animal malade et pousser en culture pure.

iii) La culture doit pouvoir causer la maladie lorsqu'elle est injectée à un animal sain.

iv) L'organisme doit pouvoir être ré-isolé à partir de l'animal infecté expérimentalement.

Les travaux de Koch ne s'arrêteront pas là et en 1890 il présentera un composant isolé à partir d'extraits de glycérol de cultures liquides du bacille, la tuberculine. Injectée à des cochons d'Inde avant et après exposition à *M. tuberculosis*, ces extraits semblaient inhiber la croissance de la bactérie. Présentée comme la solution à la plus grande maladie touchant l'homme, celle-ci se révéla malheureusement inefficace. Cependant si l'espoir d'un vaccin thérapeutique disparut, la tuberculine se révéla précieuse pour le diagnostic de personnes infectées grâce au développement du test cutané à la tuberculine, dit test Mantoux, encore utilisé de nos jours (Kaufmann and Schaible 2005), permettant de détecter une infection par la tuberculose là où l'on échoue à isoler la bactérie.

A cette époque, sans traitement, on estime que 70% des patients dont les crachats se révélaient positifs décédaient dans les 10 ans, ainsi que 20% des personnes dont les expectorations étaient négatives mais dont la culture était positive. La cure en sanatorium, qui naquît au milieu du XIXème siècle, fut le premier traitement à être utilisé globalement. Son inventeur, Hermann Brehmer (1826-1889), était un jeune botaniste qui souffrait de la tuberculose et vivait en Silésie, région s'étendant entre l'Allemagne, la Pologne et la République Tchèque. Lorsque son médecin lui conseilla de s'exposer à un climat plus clément, il partit étudier la flore himalayenne. Plus tard il rentra chez lui, soigné, et commença des études de médecine. Par la suite il fonda un hôpital logeant et soignant les tuberculeux, l'idée étant de permettre aux malades de manger sainement et de respirer continuellement de l'air frais. Les experts de l'époque étaient unanimes sur les bienfaits des traitements en extérieur et cet hôpital devînt un modèle pour les sanatoriums. Peu après ceux-ci se multiplièrent en Europe et aux Etats-Unis permettant d'une part d'isoler les personnes souffrants de la tuberculose, vis-à-vis de la population générale, tout en leur apportant un mode de vie sain censé les aider à guérir. Il reste toutefois difficile d'évaluer l'efficacité des sanatoriums car il n'y a pas eu d'études comparant rigoureusement la mortalité des patients en sanatoriums par rapport à ceux qui sont restés dans leur résidence ou ville. D'autres méthodes controversées et considérées comme dangereuses étaient parfois employées, après la découverte par Carlo Forlanini (1847-1918). La réduction du poumon infecté et sa mise au repos pouvait permettre, selon lui, un rétablissement du patient. Cette opération était effectuée par chirurgie ou bien en remplissant le poumon de gaz, le rendant non fonctionnel (pneumothorax) mais stoppant l'avancée de l'infection. Néanmoins ces méthodes furent de moins en moins utilisées et les sanatoriums fermèrent les uns après les autres après le développement de thérapies actives à base d'antibiotiques dès 1944. L'incidence de la

tuberculose continua de décroître dans les années 60, si bien que les agences gouvernementales déclarèrent qu'à l'horizon 2000, la tuberculose ne serait plus un problème de santé majeur (Spence et al. 1993).

Si la tuberculose est présente à des taux historiquement bas en Europe et en Amérique du Nord aujourd'hui, touchant principalement les personnes âgées, l'épidémie de tuberculose continue avec la même vigueur dans les pays en développement, en Asie et en Afrique subsaharienne principalement, ayant même gagné du terrain grâce à l'épidémie de VIH sévissant depuis les années 80. Il y a aujourd'hui plus de tuberculose qu'il n'y en a jamais eu, en nombre absolu, suite à l'explosion démographique humaine. De maladie touchant et tuant dans toutes les couches sociales, la tuberculose est désormais une maladie associée à la pauvreté (Spence et al. 1993). Aujourd'hui un tiers de la population mondiale est touché par ce fléau (en phase latente majoritairement) et le nouvel objectif fixé par les Nations Unies est d'éradiquer la tuberculose à l'horizon 2030 comme indiqué dans l'un des 13 « Sustainable Development Goal » (WHO 2016).

Impact sociétal de la tuberculose

La tuberculose a donc accompagné l'homme depuis la nuit des temps, tuant et façonnant l'histoire de l'humanité sûrement plus que n'importe quelle autre maladie. De nombreux témoignages et écrits montrent que la tuberculose a touché des empereurs, des rois, des reines, des poètes, des écrivains et des peintres, influençant l'histoire ainsi que l'art. Pour certains personnages il est difficile d'affirmer avec certitude qu'ils souffraient bien de la tuberculose à une époque où la bactériologie n'existait pas, cependant nous pouvons citer certains individus marquants soupçonnés d'en souffrir. L'empereur romain Hadrien ainsi que Lucius sont soupçonnés d'avoir soufferts de la tuberculose, rendant le premier incapable de voyager et ses douleurs changeant son caractère le rendant exécrable. Le second, censé lui succéder sera emporté par la maladie avant même de monter sur le trône. Le roi d'Angleterre Edward VI (1537-1553) souffrait vraisemblablement de la tuberculose également, sa mort précipita sa sœur Marie sur le trône. Si Edward n'avait pas contracté la tuberculose et n'était pas mort si jeune, le peuple anglais aurait évité les répressions brutales de « Marie la sanglante » (Chalke 1962). De nombreux autres personnages qui ont ou auraient pu marquer l'histoire sont soupçonnés d'avoir eu la phtisie, comme Madame de Pompadour, le fils de Napoléon ou Gavrillo Princip l'assassin de l'archiduc Franz Ferdinand. Au cours du XVIII^{ème} siècle la tuberculose était vue comme du vampirisme dû aux symptômes rendant les souffrants

pâles, émaciés, sensibles à la lumière et les faisant cracher du sang mais aussi de par sa transmissibilité entre proches. Plus tard, au XIX^{ème} siècle, le regard sur la tuberculose changea radicalement et les artistes de l'époque commencèrent à romantiser la maladie (Carlos et al. 2007; Chalke 1962; Daniel 2006). Les standards de beauté s'en trouvèrent modifier, et les tuberculeux aux visages pâles devinrent attirants. De très nombreux artistes ont soufferts d'une santé précaire et de la tuberculose, véritable épée de Damoclès qui a stimulé leur créativité et leur production. On peut citer de nombreux écrivains comme Orwell, Edgar Allan Poe, Balzac ou Molière. Quelques peintres également, comme Watteau par exemple, qui mourut de la tuberculose à 36 ans et dont on retrouve l'influence dans ses tableaux. Modigliani souffrait également de la tuberculose. Menant une vie dissolue d'alcoolique, il fût emporté par une méningite tuberculeuse à 35 ans. Certains philosophes comme Voltaire, Descartes ou Locke, pour ne citer qu'eux, étaient de même infectés par le bacille de Koch. Bien entendu la tuberculose n'a pas seulement influencé ceux qui en souffraient mais également tous ceux dont les proches ont été consumés.

A ce titre, le spectre de la tuberculose a fortement impressionné la société civile européenne au XIX^{ème} siècle. Cela s'est notamment traduit par une importante production de spectacles et de romans relatant le destin d'héroïnes touchées par *M. tuberculosis*. Qui ne se souvient pas de Mimi dans la Bohême de Puccini, de Hans Carstop dans la Montagne magique de Thomas Mann ou de la courtisane Marguerite Gautier dans la Dame aux Camélias de Dumas fils ?

Métriques épidémiologiques

Afin de quantifier les menaces que posent la tuberculose actuellement, les trois données principales utilisées sont l'incidence, la prévalence et la mortalité. La première indique le nombre de nouveaux cas et rechutes sur une période de temps, généralement 1 an. La prévalence mesure le nombre de cas totaux de tuberculose à un moment donné. Et la mortalité, indique le nombre de décès causés par la maladie sur une période de temps, souvent sur une année également. L'incidence ne pouvant être mesurée directement à l'échelle nationale pour des problèmes financiers et logistiques évidents, celle-ci est estimée de 4 façons différentes selon les pays et les capacités de leur système de santé et de surveillance. La méthode la plus efficace consiste à estimer l'incidence directement grâce à la notification des cas et aux diagnostiques tout en étant légèrement corrigée en prenant en compte les erreurs de diagnostiques ou les cas non reportés. Cette méthode est utilisée dans presque tous les pays « riches », la France par exemple (Figure 3) et dans certains pays en développement

notifiant les cas efficacement (Le Brésil et la Chine par exemple). Peu de pays utilisent des méthodes d'inventaire et de capture-recapture (Le Royaume-Uni par exemple). Une autre méthode consiste à prendre en compte les cas notifiés en combinaison avec l'avis d'experts (utilisée principalement dans les pays africains). La dernière méthode, qui a été utilisé pour recenser la plus grande part de l'incidence globale (62% en 2015), prend en compte les résultats d'études sur la prévalence de la maladie et sa durée pour en déduire l'incidence nationale (utilisée en Inde).

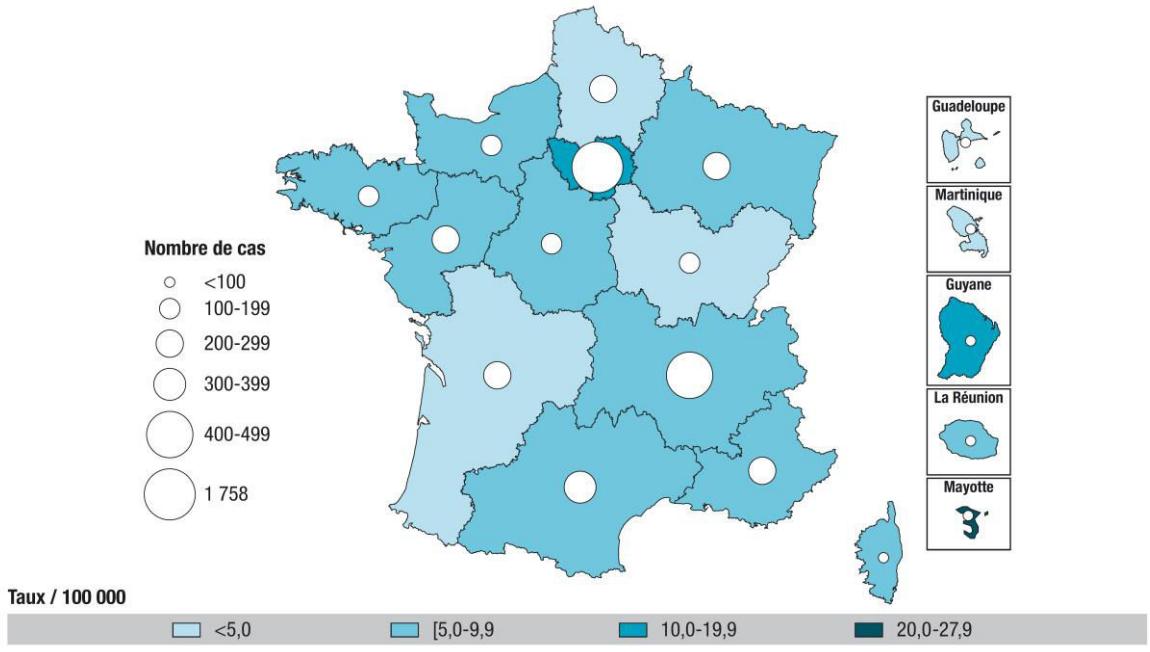


Figure 3 : Nombre de cas déclarés et taux de déclaration de tuberculose (pour 100000) par région de déclaration, France entière, 2015 (n=4741), extrait de (Guthmann et al. 2017).

Diagnostique de la maladie

Pour reporter les cas de tuberculose, il faut dans un premier temps les diagnostiquer. Cela se fait en deux temps, premièrement cliniquement, c'est-à-dire repérer les personnes atteintes de tuberculose au vu de leurs symptômes. Puis dans un deuxième temps confirmer bactériologiquement que le germe responsable de la maladie est bien *M. tuberculosis.* Or cette étape, plus compliquée à mettre en place, n'est pas systématiquement effectuée. En 2015, 57% des cas de tuberculose pulmonaire reportés à l'OMS ont été bactériologiquement confirmés. Les symptômes de la tuberculose pulmonaire sont nombreux et peuvent être présents sous diverses combinaisons. Les principaux sont la toux souvent grasse, des crachats (appelés « sputums ») souvent purulents, une hémoptysie: le fait de cracher du sang, une perte de souffle progressive, une perte de poids et de l'anorexie, de la fièvre pouvant être

accompagnée de suées nocturnes, des malaises et enfin une cachexie (dénutrition importante, fatigue et atrophie musculaire) dans les stades avancées de la maladie (Campbell and Bah-Sow 2006). Cependant, dans les pays développés, la présence de tous ces symptômes chez un patient est rare et lorsque c'est le cas, plus souvent provoqués par un cancer des poumons que par une tuberculose pulmonaire. Malheureusement dans les pays en développement, des patients présentant ces symptômes peuvent être rencontrés. Afin de diagnostiquer rapidement des cas de tuberculose, certains facteurs épidémiologiques doivent être connus. En effet, dans les pays développés, la tuberculose est beaucoup plus fréquente chez les migrants en provenance d'Afrique sub-saharienne, d'Inde, d'Asie du Sud-Est et de l'ancien bloc soviétique que chez les natifs du pays en question. Chez ces derniers, la tuberculose est diagnostiquée plus souvent chez des personnes âgées et isolées. Le fait que les personnes soient alcooliques, sans domicile fixe, consommateurs de drogues dures ou des patients immuno-déficients sont des facteurs de risque supplémentaires. Dans le reste du monde, la tuberculose touche plus communément les personnes pauvres, en état de malnutrition et/ou atteintes du VIH. On peut utiliser le test cutané à la tuberculine afin de détecter les personnes en phase latente de tuberculose, donc sans symptômes. Cependant 10 à 25% des personnes développant une tuberculose pulmonaire auront un test négatif. De plus, le test peut être positif pour des personnes vaccinés au BCG ou infecté par d'autres mycobactéries. Si certains des symptômes sont présents, avec une toux persistant plus de 3 semaines, il faut tester bactériologiquement s'il s'agit de la *M. tuberculosis*. Pour cela, 3 principales méthodes sont employées. La plus simple et la plus ancienne consiste à examiner le frottis des crachats au microscope afin d'identifier si la bactérie est présente ou non. L'examen doit se faire sur au moins deux crachats, dont un fait tôt le matin. Néanmoins, l'absence du bacille dans les crachats n'exclut pas pour autant que le patient souffre de tuberculose. La seconde est une méthode de culture après prélèvements (Figure 4). *M. tuberculosis* est un bacille à croissance très lente (3 à 6 semaines) exigeant des milieux spéciaux. Le milieu solide le plus utilisé est celui de Lowenstein-Jensen ou une de ses multiples variantes (Coletsos). Ce sont des milieux solides à base d'œufs, additionnés en proportion variable d'asparagine, de glycérine ou de vert malachite. La culture est aussi possible en milieu liquide (Middlebrook, Mycobacteria Growth Indicator Tube). Les colonies blanc-ivoire sont rugueuses et adhérentes au milieu. Elles grossissent lentement pour atteindre 3-4 mm après 2 à 3 mois. Elles ont alors un aspect en chou-fleur. La troisième méthode repose sur l'emploi de tests moléculaires rapides. Le seul actuellement recommandé par l'OMS est le Xpert® MTB/RIF assay (Cepheid, Sunnyvale USA) qui permet de tester la présence de *M. tuberculosis* et également de tester la résistance

de la souche à la rifampicine. Actuellement pour tester la résistance aux antibiotiques des souches, les méthodes cultures-dépendantes restent le golden standard. Cependant de nouveaux outils sont recommandés par l'OMS afin de permettre un meilleur diagnostique des souches résistantes. Ces tests sont appelés LPAs (Line probe assays). Basés sur l'ADN, ils permettent de tester rapidement les résistances à différents antibiotiques en identifier certaines mutations connues rendant la bactérie résistantes. Il existe plusieurs LPAs selon les types de résistantes à tester (antibiotiques de première ligne, fluoroquinolones ou antibiotiques injectables de seconde ligne).

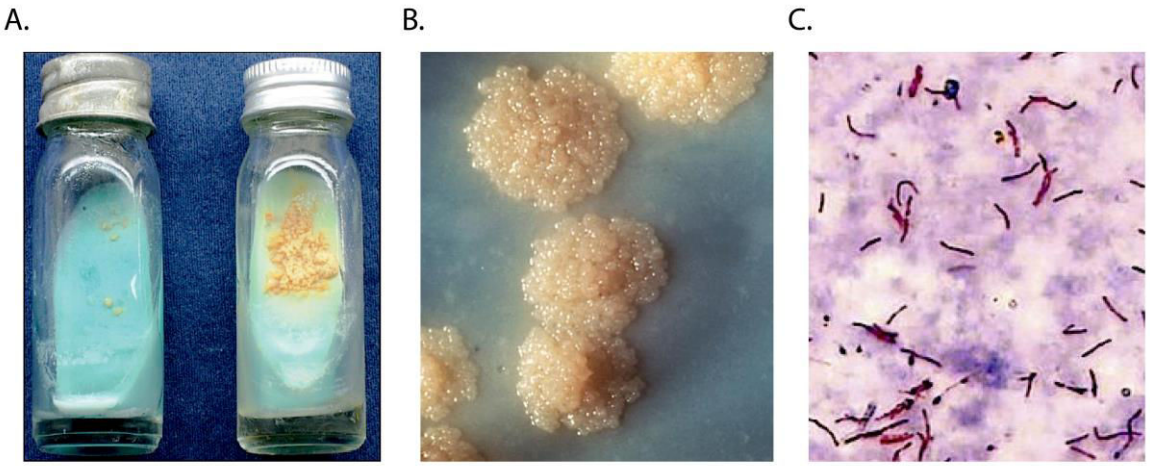


Figure 4 : a) Croissance de colonies de *M. tuberculosis* en milieu Löwenstein-Jensen (contrôle négatif et positif). b) Agrandissement de colonies typiques de type rugueuse, dites en choux-fleur. c) Coloration au Ziehl-Neelsen du bacille de Koch.

Chiffres et bilans mondiaux 2015 (WHO 2016)

Mondialement la tuberculose toucherait entre 2 et 3 milliards de personnes de façon asymptomatique, en phase dite latente. Parmi ceux-ci, seuls 5 à 15% développeront la maladie durant leur vie, avec une probabilité accrue chez les patients immuno-déficients, affaiblis par l'âge ou en état de malnutrition. En 2015, d'après le World Tuberculosis Report 2016 (OMS), les principaux chiffres de la maladie sont les suivants : Tout d'abord, on estime l'incidence à 10.4 millions de nouveaux cas (dans une fourchette de 8.4 millions à 12.2 millions) pour un total de 6.1 millions de cas notifiés et reportés à l'OMS (Figure 5). Parmi ceux-ci 5.9 millions d'hommes, 3.5 millions de femmes et 1 million d'enfants. Comme vu précédemment, certains facteurs à risque augmentent la probabilité de développer la tuberculose, ce qui explique que les hommes soient plus touchés. Les personnes atteintes du VIH représentent quant à elles 1.2 millions de cas. En 2015, la maladie a entrainé la mort de 1.4 millions de malades dont 0.4 millions souffrant, en combinaison, de la tuberculose et du VIH. Les pays les plus touchés sont des pays en voie de développement. En effet, 60% des nouveaux cas ont été déclarés

dans 6 pays: l'Inde, l'Indonésie, la Chine, le Pakistan, le Nigéria et l'Afrique du Sud. En Afrique l'épidémie de tuberculose est facilitée par l'épidémie de VIH et ainsi 31% des nouveaux cas de tuberculose y sont des patients porteurs du virus, atteignant même 50% en Afrique du Sud. La tuberculose reste donc en 2015 une des dix premières causes de décès à l'échelle mondiale, devant le VIH, malgré un déclin du nombre de nouveaux cas de 1.5% par rapport à 2014.

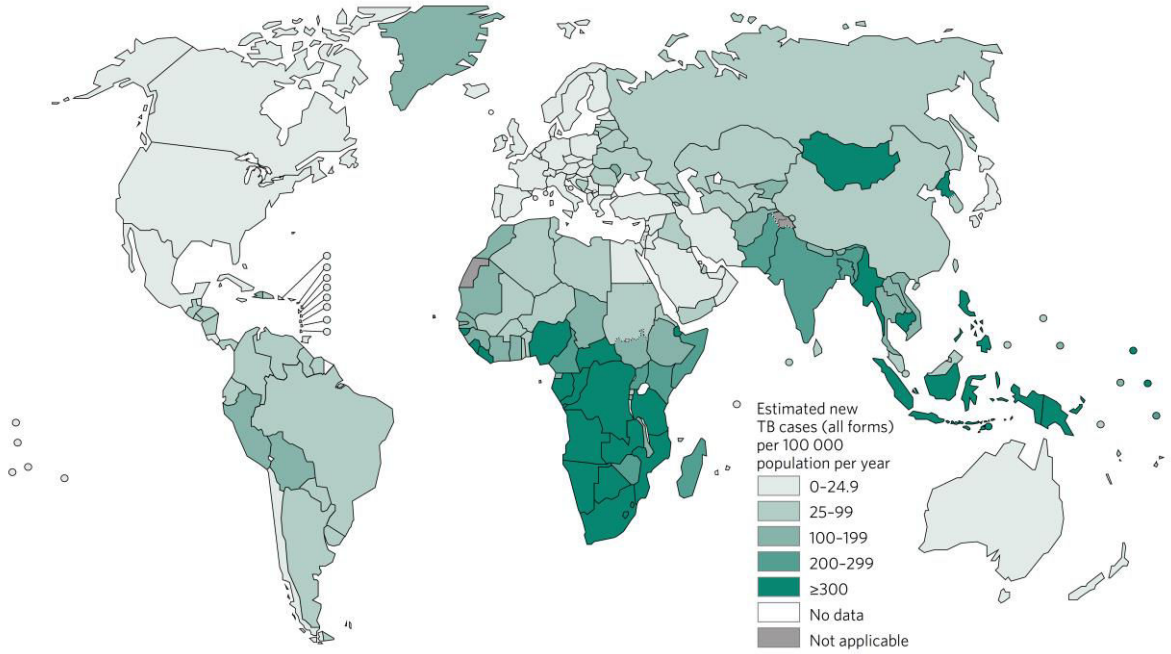


Figure 5 : Estimation de l'incidence de la tuberculose à l'échelle mondiale en 2015. Extrait du World Tuberculosis Report 2016.

Si ces résultats sont encourageants, ceux concernant les souches résistantes aux antibiotiques le sont moins. En effet, en 2015 on estime à 100000 le nombre de souches résistantes à la rifampicine et à 480000 les nouveaux cas de souches MDR (multi-drug resistant), provoquant la mort de 250000 personnes. Très mal répartis à l'échelle mondiale, 45% des cas sont réunis dans seulement 3 pays : L'Inde, la Chine et la Russie (Figure 6). L'évolution du nombre de cas résistants est difficile à évaluer car parmi les pays où ces souches sont particulièrement prévalentes les données ne sont pas enregistrées ou alors depuis peu de temps. Cependant sur les quelques pays où les données sont disponibles depuis plus de 3 ans, la proportion des souches MDR parmi les nouveaux cas a tendance à augmenter soit parce que le nombre de nouveaux cas MDR augmente soit parce qu'il diminue moins rapidement que le nombre de cas de tuberculose globaux, selon les pays en question.

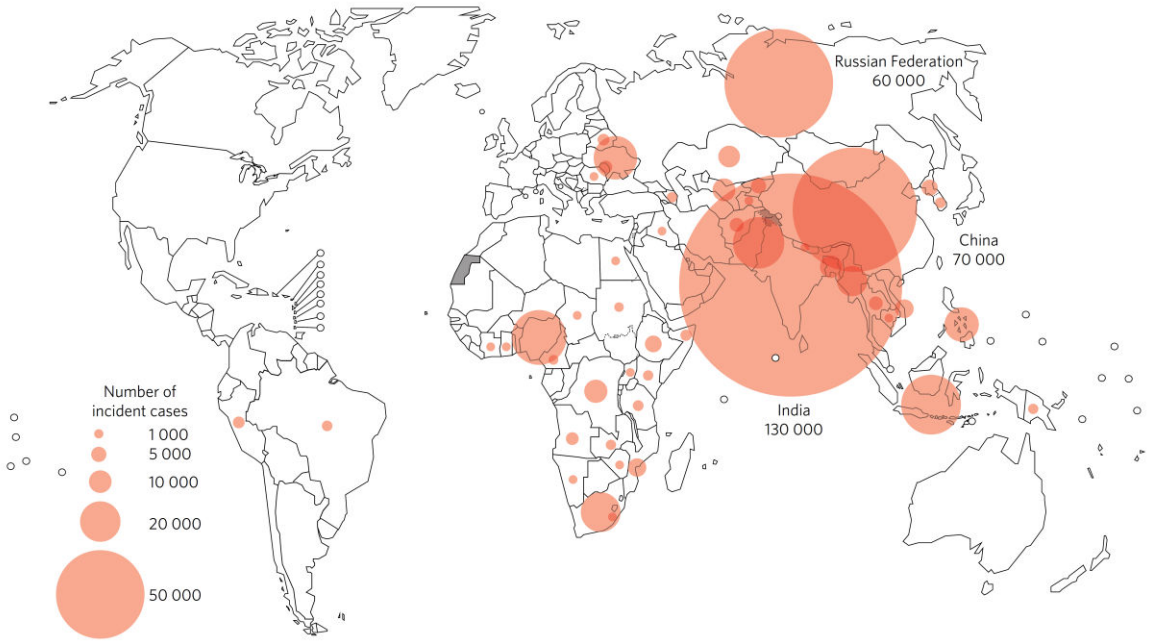


Figure 6 : Estimation de l'incidence des souches MDR et Rifampicine résistantes pour les pays avec plus de 1000 cas notifiés, en 2015. Extrait du Wolrd Tuberculosis Report 2016.

Traitements

Les premiers antituberculeux furent utilisés à partir de 1944, d'abord la streptomycine seule, puis en combinaison avec de l'acide para-aminosalicylic (PAS). Avec ce traitement, 80% des personnes souffrants de tuberculose pulmonaire récupéraient. Malheureusement, on constata rapidement que les traitements ne permettaient pas de soigner tous les patients, l'identification de résistances chez les bactéries ne tardant pas. De nouveaux antibiotiques furent découverts et développés dans les années qui suivirent, jusqu'à une vingtaine aujourd'hui, mais des résistances apparaissent pour chaque type d'antituberculeux.

La méthode la plus efficace pour guérir au mieux le patient et éviter le développement de résistances est d'utiliser plusieurs antibiotiques en combinaison. Premièrement car une personne infectée peut l'être par diverses populations ayant chacune des profils de résistance propres, mais aussi pour frapper plus efficacement la bactérie et éviter le développement de souches résistantes durant la thérapie en se basant sur une approche probabiliste. Les cas de tuberculose en phase latente peuvent également être traités, la posologie recommandée reposant sur une prise d'isoniazide pendant 9 mois (Kasner et al. 2013). Le traitement standard pour les cas de tuberculose active, des variations peuvent exister selon les pays, consiste à donner au patient de la rifampicine, de l'isoniazide, de l'ethambutol et de la pyrazinamide pendant 2 mois, puis des tests de sensibilités sont effectués et si l'organisme n'est pas résistant à la rifampicine et à l'isoniazide, ces deux antibiotiques doivent être pris

pendant encore 4 mois tandis que les deux autres sont arrêtés (Campbell and Bah-Sow 2006). Le traitement de la tuberculose est donc long, 6 mois de traitement pour une personne atteinte d'une souche non résistante. Aujourd'hui, malgré sa longueur, le traitement standard a un coût faible, entre 100 et 1000 dollars pour les 6 mois, et est relativement efficace avec 85% de réussite. Pour les 15% restants, la cause principale d'échec de la thérapie est le développement de résistances chez la souche infectant le patient. Ainsi entre 2000 et 2015, les traitements ont permis de sauver 49 millions de personnes.

Concernant les souches résistantes, les traitements sont plus lourds et les antibiotiques plus toxiques pour le patient doivent être pris sur une plus longue durée (Table 1). Certaines souches peuvent être résistantes à 1 ou 2 antibiotiques qui sont remplaçables. Cependant la situation est plus complexe dans le cas de souches MDR, définis comme résistantes à l'isoniazide et à la rifampicine; les deux antituberculeux de première ligne les plus efficaces. Jusqu'en 2016, le traitement durait 20 mois et coûtait entre 2000 et 5000 dollars par patient. Récemment, un nouveau traitement standardisé pour MDR est recommandé par l'OMS (WHO 2016), sa durée a été réduite à 9-12 mois et le coût chute à 1000 dollars. Malgré les progrès sur la durée et le coût de ce traitement, beaucoup de travail reste à faire, notamment sur son accessibilité qui doit encore être facilité et son efficacité qui doit être améliorée. En effet, seulement 125000 personnes (20% des patients éligibles) ont pu profiter du traitement avec un taux de succès de 52%. Au sein des souches MDR, 9,5% accumulent encore plus de résistances, appelées en conséquence XDR (extensively drug resistant) elles sont résistantes à au moins une fluoroquinolone et un agent injectable de seconde ligne, les médicaments utilisés dans le cas des MDR. Chez les patients infectés par les souches XDR on observe donc un taux de réussite encore plus faible, seulement 28%. Afin de lutter contre ces cas extrêmement préoccupants, deux nouveaux antibiotiques ont été introduits récemment, la bedaquiline et la delamanide; d'autres sont actuellement en développement (D'Ambrosio et al. 2015).

Groupe	Antibiotiques
1) Antibiotiques anti-tuberculeux oraux de 1ère ligne	Isoniazide (H)
	Rifampicin (R)
	Ethambutol (E)
	Pyrazinamide (Z)
	Rifabutine (Rfb)
	Rifapentine (Rpt)
2) Antibiotiques anti-tuberculeux injectables	Streptomycine (S)
	Kanamycine (Km)
	Amikacine (Am)
	Capréomycine (Cm)
3) Fluoroquinolones	Moxifloxacine (Mfx)
	Lévofloxacine (Lfx)
	Gatifloxacine (Gfx)
	Ofloxacine (Ofx)
	Ciprofloxacine (Cip)
4) Antibiotiques anti-tuberculeux oraux de 2nd ligne	Ethionamide (Eto)
	Prothionamide (Pto)
	Cyclosérine (Cs)
	Térizidone (Trd)
	Para-aminosalicylic acid (PAS)
5) Anti-tuberculeux avec peu d'informations sur l'efficacité et/ou la dangerosité à long terme dans le traitement de souche résistantes (inclus les nouveaux agents)	Bédaquiline (Bdq)
	Délamanide (Dlm)
	Linézolide (Lzd)
	Clofazimine (Cfz)
	Amoxicilline/Clavulanate (Amx/Clv)
	Imipénème/Cilastatine Ipm/Cln)
	Méropénème (Mpm)
	Haute dose d'isoniazide
	Thioacétazone (T)
	Clarithromycine (Clr)

Table 1 : Liste des agents anti tuberculeux et leur classification (World Health Organization (WHO) 2014).

Stratégies vaccinales

Le seul vaccin contre la tuberculose, le BCG (Bacille Calmin-Guérette) a été développé en 1921 et est toujours utilisé aujourd'hui. Ce vaccin est une souche atténuée de *M. bovis* obtenue par repiquage, 230 fois entre 1908 et 1921. Le BCG est le vaccin le plus utilisé dans le monde, généralement donné aux nouveaux nés ou aux enfants et est actuellement administré dans plus de 167 pays et est obligatoire dans au moins 64 d'entre eux. Il est recommandé par l'OMS dans tous les pays présentant une forte prévalence car il permet de protéger contre les formes sévères de tuberculose chez l'enfant, souvent mortelles, comme la

méningite tuberculeuse ou la tuberculose miliaire. Rares sont les pays qui n'ont pas de programme de vaccination au BCG, ceux-ci comprennent les Etats-Unis, le Canada, l'Italie, la Belgique et les Pays-Bas.

Toutefois, son efficacité chez l'adulte est largement débattue. En effet, son niveau de protection contre la tuberculose pulmonaire semble très variable entre populations (Andersen and Doherty 2005). Cependant, certaines études semblent tout de même bien montrer son efficacité (Michelsen et al. 2014; Roy et al. 2014). Fait intéressant, le vaccin BCG confère une immunité à la lèpre, causée par une mycobactérie également, *Mycobacterium leprae.* Dans le but d'éradiquer la maladie, développer un nouveau vaccin est essentiel. Aujourd'hui 13 vaccins sont actuellement en développement, à différents stades, le but étant d'obtenir un vaccin qui protégerait contre une préexposition, mais aussi qui préviendrait du développement de la maladie chez les personnes atteintes en phase latente, qui représentent un tiers de la population mondiale (Ahsan 2015).

De plus, un des nombreux problèmes étant les co-infections VIH/TB, l'innocuité du vaccin est donc cardinale. Celui-ci devrait donc être au moins aussi atténué que le BCG mais plus efficace. Enfin, une très grande proportion de la population étant vacciné par le BCG, il faudra que le vaccin ne présente pas d'interactions néfastes avec ce dernier. La situation est complexe, d'autant plus que les réinfections par de nouvelles souches après traitement sont fréquentes et donc une primo infection ne permet pas développer une immunité totale chez le patient soigné (van Rie et al. 1999). Le vaccin doit donc être plus efficace qu'une infection de *M. tuberculosis*.

***M. tuberculosis* à l'ère de la génomique - Objectifs**

Le génome de la souche la mieux connue de *M. tuberculosis*, H37Rv, fut séquencé pour la première fois en 1998 (Cole et al. 1998). C'est un chromosome circulaire unique de 4,4 millions de paires de bases à fort taux en GC (65.6%) et qui comporte environ 4000 gènes. La bactérie se développant dans les poumons de l'hôte, donc ne rencontrant qu'exceptionnellement des clones génétiquement distincts, la contribution relative de la recombinaison demeure très modeste. *M. tuberculosis* évolue essentiellement par mutations, sous l'influence de la sélection et de la dérive génétique, depuis le goulet d'étranglement correspondant au franchissement du statut d'espèce environnementale à celui d'agent pathogène obligatoire de l'homme (Achtman 2008; Comas and Gagneux 2011; Pepperell et al. 2013; Supply et al. 2013). Avec un taux de mutation modeste (Ford et al. 2011), le MTBC

présente une structure de population extrêmement clonale et une faible diversité génétique. D'autres études semblent montrer la présence de recombinaison chez *M. tuberculosis*, au moins dans certaines parties de son génome (Liu et al. 2006; Namouchi et al. 2012; Phelan et al. 2016). Ces transferts horizontaux pourraient avoir lieu lors d'infections multiples (un patient infecté par deux clones génétiquement distincts pendant la même période) qui atteindraient jusqu'à 20% des cas dans certaines conditions particulières, mais qui restent néanmoins difficiles à tracer.

Avec le développement de la biologie moléculaire un grand nombre de techniques de génotypage ont été développées, permettant de grandes avancées en épidémiologie. Dans le cas de la tuberculose, la première méthode de génotypage ayant été utilisée et appliquée dans des études d'épidémiologie moléculaire est l'IS6110 DNA fingerprinting (Van Embden et al. 1993). D'autres types de fingerprint ont également été développés mais rapidement ces méthodes très chronophages, modérément reproductibles et relativement chères furent remplacées par d'autres, basées sur la PCR, beaucoup moins lourdes et plus économiques (Niemann and Supply 2014). Les deux fers de lance de ces techniques sont le spoligotyping, basé sur la détection d'espaceurs dans une région de répétitions directes (région CRISP-R), censées être présentes chez toutes les souches (Kamerbeek et al. 1997) : L'absence ou la présence de ces séquences d'espacement en diverses combinaisons permet de discriminer les souches les unes des autres. Et la seconde, le typage MIRU-VNTR, basé sur le typage de loci, principalement intergéniques, consistant en un nombre variable de répétitions en tandem, à la manière des mini-satellites (Supply et al. 1997). Ces méthodes ont permis des avancées considérables dans la compréhension de l'évolution de la *M. tuberculosis*, cependant ils n'interrogent qu'une partie infime du génome bactérien (Niemann and Supply 2014). Ainsi avec l'apparition des méthodes NGS (Next generation sequencing) permettant d'avoir accès peu ou prou au génome complet des bactéries, l'épidémiologie moléculaire par séquençage de génomes complets (WGS) a connu un véritable essor. La masse d'information apportée par cette génération de données permet d'adresser de nouvelles questions et sujets à explorer tels que la démographie de *M. tuberculosis*, son âge ainsi que son origine, l'accès à des mutations ciblant des gènes précis et donc à leur fonction. Cette thèse s'inscrit dans cette dynamique, elle s'appuie sur les nouvelles technologies de séquençage et le typage MIRU-VNTR. Mais elle s'adosse aussi sur de grands ensembles (gros jeu de données) afin d'éviter l'écueil de l'anecdotique et de mieux appréhender les mécanismes évolutifs en marche (puissance statistique et phénomènes à large échelle). Au cours du chapitre 1, nous reviendrons plus en

détail sur l'apport récent des données génétiques et des NGS dans la compréhension de l'évolution du MTBC ainsi que les changements de paradigmes au cours de sa découverte. Une fois tout le contexte posé nous aborderons les questions posées et les moyens mis en place pour y répondre.

Comme mentionné plus haut, la menace des souches résistantes prend de plus en plus d'ampleur et pose de sérieux problèmes de santé publique. Or les mécanismes permettant aux souches d'accumuler un grand nombre de mutations conférant des résistantes sont encore mal connus. Certaines lignées semblent acquérir des résistances plus rapidement que d'autres (Ford et al. 2013) sans que les raisons soient réellement connues. Notamment la lignée Beijing (lignée 2), un clone de *M. tuberculosis* originaire d'Asie, particulièrement virulent et comportant une grande proportion de souches résistantes (Merker et al. 2015; Zhang et al. 2013). Sans nul doute que des facteurs externes tels que la politique de santé et le PIB des pays entrent en compte dans le développement de souches résistantes. Néanmoins, une fraction non négligeable de cette composante repose sur des facteurs génétiques propres à la bactérie. Effectivement, les premières épidémies de souches hautement résistantes touchaient principalement des patients souffrant du VIH (Frieden et al. 1996) ; on pensait donc naturellement que les résistances de ces souches s'accompagneraient d'un coût qui réduirait leur fitness et transmissibilité, empêchant ainsi la prolifération des germes MDR et XDR (Andersson and Levin 1999). Cependant, certaines bactéries semblent désormais hautement résistantes tout en étant très infectieuses (Comas et al. 2011). Ce phénomène serait dû à l'acquisition de mutations dites compensatoires par ces bactéries, comme leur nom l'indique, compensant le coût de fitness associé aux résistances (Li et al. 2016; De Vos et al. 2013; Handel et al. 2006; Meftahi et al. 2015). Nous traiterons ces questions en Chapitre 2 et 3 en nous intéressant tout d'abord au cas des souches résistantes en Ouzbékistan, dans la région de Nukus, où nous nous sommes intéressés aux possibles événements ayant favorisées l'installation de ces souches. Puis plus spécifiquement en étudiant le clone W148, faisant parti de la lignée Beijing et composé uniquement de souches MDR provenant de l'ex-URSS et présentes dans toute l'Europe. Notre apport principal est l'établissement d'un listing de mutations compensatoires, ainsi que l'identification de SNPs (Single nucleotide polymorphisms) associés à l'acquisition de nouvelles résistances aux antibiotiques.

Au cours du chapitre 4 je me suis penché sur l'adaptation de *M. tuberculosis* à toute sa diversité d'hôtes. Comme mentionné plus haut, la tuberculose touche toute une variété de mammifères allant des bovins aux pinnipèdes en passant par les primates. L'homme

s'intéressant naturellement plus à *H. sapiens* et ses pathogènes, la coévolution entre *M. tuberculosis* et *H. sapiens* (Berg and Smith 2014; Behr and Gordon 2015) est bien mieux connue que celle des souches animales à leurs hôtes respectifs. *M. bovis* est relativement bien connue car elle peut être transmise à l'homme s'il boit du lait non pasteurisé provenant de vaches infectées, de plus le BCG en est une souche dérivée. Mais en dehors de rares épisodes, lorsque *M. tuberculosis* est transmise à un animal, elle ne provoque pas d'épidémie. Réciproquement les souches animales ne sont que très rarement transmises d'hommes à hommes (Berg and Smith 2014; Behr and Gordon 2015; Bos et al. 2014). Les différentes lignées du MTBC ne divergent que de l'ordre de 0.05% entre elles, pourtant elles semblent toutes réellement adaptées à leur hôte. Vraisemblablement la *M. tuberculosis*, pathogène de l'homme, son hôte primaire, est passée secondairement aux bovins et aux autres mammifères (Gibbons 2008). Il ne s'agit donc pas d'une zoonose mais plutôt du contraire. Nous nous sommes intéressés aux adaptations ayant permis ces changements d'hôte successifs, en cherchant des mutations au sein de gènes possiblement sous sélection. Une piste avait déjà été explorée avec des mutations découvertes au sein du régulon PhoP/PhoR, qui entraineraient une baisse de virulence chez l'homme (Gonzalo-Asensio et al. 2014) et seraient présentes chez les souches animales proches.

Enfin, nous nous sommes intéressés au « succès » individuel des clones de *M. tuberculosis* en développant de nouveaux outils statistiques et indices (Chapitres 5 et 6). Dans un cadre large, lorsqu'on parle d'un clone ayant du succès, il s'agit en fait de celui de sa lignée. Cet avantage se mesure par sa transmissibilité, sa virulence, sa fitness ou une combinaison subjective de toutes ces caractéristiques. Ainsi, la lignée Beijing est jugée comme ayant un avantage sélectif car elle comporte un grand nombre de souches résistantes, s'est répandue mondialement en relativement peu de temps et a vu sa population augmenter significativement au cours du dernier siècle (Merker et al. 2015; Luo et al. 2015). Ainsi on pourra argumenter que certaines mutations uniquement présentes dans la lignée Beijing, ou la caractérisant, contribuent à son succès, mais cela est basé sur une mesure qualitative du succès. De plus cela n'est pas souche spécifique mais se base sur le succès qu'on assigne à un clade. Pour mesurer plus quantitativement le succès des souches, nous avons en chapitre 2 et 3 utilisé un indice de transmission calculé pour chaque individu. Cet indice est la mesure du nombre de souches distantes de moins de 10 SNPs pour chacune d'elles. Cela permet donc d'avoir une mesure quantitative, souche spécifique, pouvant être corrélée avec différentes mesures et informations. Cependant le fait de fixer une limite, de 10 SNPs dans ce cas, et

compter le nombre de souches proches a un côté arbitraire qui n'est pas satisfaisant. Afin de palier à cela, nous avons développé un indice souche spécifique permettant de mesurer le succès de façon temps dépendante en introduisant les termes de succès épidémique et endémique. Nous avons appliqué cette méthode sur les données génotypiques obtenues sur 1641 patients atteints de tuberculose puis avons pu corréler cet indice avec des informations épidémiologiques, cliniques et génétiques dans le but d'estimer les paramètres influençant le succès des souches.

Bibliographie

Achtman M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* **62**: 53–70.

Ahsan MJ. 2015. Recent advances in the development of vaccines for tuberculosis. *Ther Adv Vaccines* **3**: 66–75.

Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, van Pittius NCG. 2010. Novel Mycobacterium tuberculosis complex pathogen, M. Mungi. *Emerg Infect Dis* **16**: 1296–1299.

Andersen P, Doherty TM. 2005. Opinion: The success and failure of BCG — implications for a novel tuberculosis vaccine. *Nat Rev Microbiol* **3**: 656–662.

Andersson DI, Levin BR. 1999. The biological cost of antibiotic resistance. *Curr Opin Microbiol* **2**: 489–493.

Aranaz A, Cousins D, Mateos A, Domínguez L. 2003. Elevation of Mycobacterium tuberculosis subsp. caprae Aranaz et al. 1999 to species rank as Mycobacterium caprae comb. nov., sp. nov. *Int J Syst Evol Microbiol* **53**: 1785–1789.

Bates JH, Stead WW. 1993. The history of tuberculosis as a global epidemic. *Med Clin North Am* **77**: 1205–1217.

Behr M a, Gordon S V. 2015. Why doesn ' t Mycobacterium tuberculosis spread in animals ? *Trends Microbiol* **23**: 1–2.

Berg S, Smith NH. 2014. Why doesn't bovine tuberculosis transmit between humans? *Trends Microbiol* **22**: 552–553.

Boniotti MB, Gaffuri A, Gelmetti D, Tagliabue S, Chiari M, Spisani M, Nassuato C, Gibelli L, Sacchi C, Zanoni M, et al. 2014. Detection and molecular characterization of Mycobacterium microti in wild boar from northern Italy. *J Clin Microbiol* **52**: 2834–2843.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest S a., Bryant JM, Harris SR, Schuenemann VJ, et al. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*.

Campbell IA, Bah-Sow O. 2006. Pulmonary tuberculosis: diagnosis and treatment. *BMJ* **332**: 1194–1197.

Carlos PJ, Sylvia Cardoso Leao, Ritacco V. 2007. Tuberculosis 2007. 687.

Chalke HD. 1962. the Impact of Tuberculosis on History, Literature and Art. *Med Hist* **6**: 301–318.

Chisholm RH, Trauer JM, Curnoe D, Tanaka MM. 2016. Controlled fire use in early humans might have triggered the evolutionary emergence of tuberculosis. *Proc Natl Acad Sci* **113**: 9051–9056.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V., Eiglmeier K, Gas S, Barry CE, et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**: 537–544.

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S. 2011. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* **44**: 106–110.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**: 1176–1182.

Comas I, Gagneux S. 2011. A role for systems epidemiology in tuberculosis research. *Trends Microbiol* **19**: 492–500.

Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, Dabrowski PW, Radonic A, Niemann S, Parkhill J, et al. 2013. Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. *Emerg Infect Dis* **19**: 969–976.

Cousins D V., Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, Duignan P, Murray A, Dupont C, Ahmed N, et al. 2003. Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *Int J Syst Evol Microbiol* **53**: 1305–1314.

D'Ambrosio L, Centis R, Sotgiu G, Pontali E, Spanevello A, Migliori GB. 2015. New anti-tuberculosis drugs and regimens: 2015 update. *ERJ Open Res* **1**: 00010–02015.

Daniel TM. 2006. The history of tuberculosis. *Respir Med* **100**: 1862–1870.

De Vos M, Müller B, Borrell S, Black PA, Van Helden PD, Warren RM, Gagneux S, Victor TC. 2013. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob Agents Chemother* **57**: 827–832.

Delogu G, Sali M, Fadda G. 2013. The biology of mycobacterium tuberculosis infection. *Mediterr J Hematol Infect Dis* **5**.

Dippenaar A, Parsons SDC, Sampson SL, van der Merwe RG, Drewe JA, Abdallah AM, Siame KK, Gey van Pittius NC, van Helden PD, Pain A, et al. 2015. Whole genome sequence analysis of Mycobacterium suricattae. *Tuberculosis* **95**: 682–688.

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, et al. 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* **43**: 482–486.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* **45**: 784–790.

Frieden TR, Sherman LF, Maw KL, Fujiwara PI, Crawford JT, Nivin B, Sharp V, Hewlett D, Brudney K, Alland D, et al. 1996. A multi-institutional outbreak of highly drug-resistant tuberculosis: epidemiology and clinical outcomes. *JAMA* **276**: 1229–1235.

Gengenbacher M, Kaufmann SHE. 2012. Mycobacterium tuberculosis: Success through dormancy. *FEMS Microbiol Rev* **36**: 514–532.

Gibbons. 2008. Tuberculosis Jumped From Humans to Cows , Not Vice Versa. *Science (80-)* **320**: 608.

Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, Laval F, Daffé M, Martin C, Brosch R, et al. 2014. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A* **111**: 11491–11496.

Guthmann JP, Levy Bruhl D, Ait Belghitti F. 2017. Epidémiologie de la tuberculose en France en 2015. Impact de la suspension de l'obligation vaccinale BCG sur la tuberculose de l'enfant, 2007-2015. *Numéro thématique Journée Mond lutte contre la Tuberc 24 mars 2017* 116–126.

Handel A, Regoes RR, Antia R. 2006. The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput Biol* **2**: 1262–1270.

Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, Galili E, Eshed V, Greenblatt CL, Lemma E, et al. 2008. Detection and Molecular Characterization of 9000-Year-Old Mycobacterium tuberculosis from a Neolithic Settlement in the Eastern Mediterranean. *PLoS One* **3**: e3426.

Hoffmann C, Leis A, Niederweis M, Plitzko JM, Engelhardt H. 2008. Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. *Proc Natl Acad Sci U S A* **105**: 3963–3967.

Kamerbeek J, Schouls L, Kolk A, Van Agterveld M, Van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, et al. 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol* **35**: 907–914.

Karlson AG. 1970. Mycobacterium Bovis Nom. Nov. *Int J Syst Bacteriol* **20**: 273–282.

Kasner E, Hunter CA, Ph D, Kariko K, Ph D. 2013. NIH Public Access. *J Immunol* **70**: 646–656.

Kaufmann SHE, Schaible UE. 2005. 100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus. *Trends Microbiol* **13**: 469–475.

Koch R. 1882. Die Aetiologie der Tuberculose (Nach einem in der physiologischen Gesellschaft zu Berlin am 24. März gehaltenem Vortrage). *Berliner klin Wochenschr* **19**: 221–30.

Kumar S V, Deka MK, Bagga M, Kala MS, Gauthaman K. 2010. A systematic review of different type of tuberculosis. *Eur Rev Med Pharmacol Sci* **14**: 831–843.

Li QJ, Jiao WW, Yin QQ, Xu F, Li JQ, Sun L, Xiao J, Li YJ, Mokrousov I, Huang HR, et al. 2016. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant Mycobacterium tuberculosis Beijing genotype strains in China. *Antimicrob Agents Chemother* **60**: 2807–2812.

Liu X, Gutacker MM, Musser JM, Fu YX. 2006. Evidence for recombination in Mycobacterium tuberculosis. *J Bacteriol* **188**: 8169–8177.

Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G, et al. 2015. Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. *Proc Natl Acad Sci* **112**: 8136–8141.

Meftahi N, Namouchi A, Mhenni B, Brandis G, Hughes D, Mardassi H. 2015. Evidence for the critical role of a secondary site *rpoB* mutation in the compensatory evolution and successful transmission of an MDR tuberculosis outbreak strain. *J Antimicrob Chemother* dkv345.

Merker M, Blin C, Mona S, Duforet-frebourg N, Lecher S, Willery E, Blum M, Rüsch-gerdes S, Mokrousov I, Aleksic E, et al. 2015. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat Genet*.

Michelsen SW, Soborg B, Koch A, Carstensen L, Hoff ST, Agger EM, Lillebaek T, Sorensen HCF, Wohlfahrt J, Melbye M. 2014. The effectiveness of BCG vaccination in preventing *Mycobacterium tuberculosis* infection and disease in Greenland. *Thorax* **69**: 851–856.

Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. 2012. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Res* **22**: 721–734.

Niemann S, Supply P. 2014. Diversity and evolution of Mycobacterium tuberculosis: moving to whole-genome-based approaches. *Cold Spring Harb Perspect Med* **4**: a021188.

Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. *PLoS Pathog* **9**: e1003543.

Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC, Glynn JR, Crampin AC, Alves A, et al. 2016. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17**: 151.

Pigrau-Serrallach C, Rodriguez-Pardo D. 2013. Bone and joint tuberculosis. *Eur Spine J* 556–566.

Rodriguez-Campos S, Smith NH, Boniotti MB, Aranaz A. 2014. Overview and phylogeny of Mycobacterium tuberculosis complex organisms: Implications for diagnostics and legislation of bovine tuberculosis. *Res Vet Sci* **97**: S5–S19.

Rothschild B, Martin L, Lev G, Bercovier H, Kahila Bar-Gal G, Greenblatt C, Donoghue H, Spigelman M, Brittain D. 2001. Mycobacterium tuberculosis complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis* **33**: 305–311.

Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, Snell L, Mangtani P, Adetifa I, Lalvani A, et al. 2014. Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *Bmj* **349**: g4643–g4643.

Smith I. 2003. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev* **16**: 463–496.

Spence D, Spence D, Hotchkiss J, Hotchkiss J, Williams C, Williams C, Davies P, Davies P. 1993. Tuberculosis and poverty. *Bmj* **307**: 759–761.

Supply P, Magdalena J, Himpens S, Locht C. 1997. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26**: 991–1003.

Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, et al. 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nat Genet* **45**: 172–9.

Van Embden JD a, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, et al. 1993. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: Recommendations for a standardized methodology. *J Clin Microbiol* **31**: 406–409.

van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, van Soolingen D. 2012. Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies. *Emerg Infect Dis* **18**: 653–655.

van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, Beyers N, van Helden PD. 1999. Exogenous Reinfection as a Cause of Recurrent Tuberculosis after Curative Treatment. *N Engl J Med* **341**: 1174–1179.

WHO. 2016. Global Tuberculosis Report 2016. *Cdc 2016* 214.

Winglee K, Manson McGuire A, Maiga M, Abeel T, Shea T, Desjardins CA, Diarra B, Baya B, Sanogo M, Diallo S, et al. 2016. Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. *PLoS Negl Trop Dis* **10**: e0004332.

World Health Organization (WHO). 2014. *Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis*.

Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, et al. 2013. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet* **45**: 1255–60.

Zink Albert, Haas CJ, Reischl U, Szeimies U, Nerlich AG. 2001. Molecular Analysis of Skeletal Tuberculosis in an Ancient Egyptian Population. *J Med Microbiol* **50**: 355–366.

Chapitre 1.

Histoire évolutive, démographique et migratoire du *Mycobacterium tuberculosis* complexe

Microbiology Spectrum
American Society for Microbiology Press

The Evolutionary History, Demography, and Spread of the *Mycobacterium tuberculosis* Complex

MAXIME BARBIER and THIERRY WIRTH

Laboratoire Biologie Intégrative des Populations, Evolution Moléculaire; Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Univ. Pierre et Marie Curie, EPHE, Sorbonne Universités, 75231 Paris cedex 05, France

ABSTRACT With the advent of next-generation sequencing technology, the genotyping of clinical *Mycobacterium tuberculosis* strains went through a major breakup that dramatically improved the field of molecular epidemiology but also revolutionized our deep understanding of the *M. tuberculosis* complex evolutionary history. The intricate paths of the pathogen and its human host are reflected by a common geographical origin in Africa and strong biogeographical associations that largely reflect the past migration waves out of Africa. This long coevolutionary history is cardinal for our understanding of the host-pathogen dynamic, including past and ongoing demographic components, strains' genetic background, as well as the immune system genetic architecture of the host. Coalescent- and Bayesian-based analyses allowed us to reconstruct population size changes of *M. tuberculosis* through time, to date the most recent common ancestor and the several phylogenetic lineages. This information will ultimately help us to understand the spread of the Beijing lineage, the rise of multidrug-resistant sublineages, or the fall of others in the light of socioeconomic events, antibiotic programs, or host population densities. If we leave the present and go through the looking glass, thanks to our ability to handle small degraded molecules combined with targeted capture, paleomicrobiology covering the Pleistocene era will possibly unravel lineage replacements, dig out extinct ones, and eventually ask for major revisions of the current model.

INTRODUCTION

Tuberculosis has plagued mankind over the centuries and probably accompanied modern *Homo sapiens* out of Africa. The epidemiological agent of phthisis, also known as "consumption," reached its epidemic apex during the 18th and 19th centuries. During the industrialization era, the disease was associated with the concentration of labor and poor socioeconomic settings that ultimately favored the spread of this "crowd" pathogen. This high-burden period was then followed by a progressive decline of the death and disease tolls that predated the antibiotic era and the *Mycobacterium bovis* BCG vaccination. The evolutionary histories of the host and its pathogen are intricately associated, implying that tuberculosis can only be fully understood in the light of *H. sapiens* origins, migrations, and demography (1). Excluding these parameters from our analyses might lead us to false conclusions regarding evolution, epidemiology, and pathobiology. In the same line, there is also an urgent need to unravel the genomic features

Received: 13 January 2016, **Accepted:** 21 January 2016, **Published:** 12 August 2016

Editors: William R. Jacobs Jr., Howard Hughes Medical Institute, Albert Einstein School of Medicine, Bronx, NY 10461; Helen McShane, University of Oxford, Oxford OX3 7DQ, United Kingdom; Valerie Mizrahi, University of Cape Town, Rondebosch 7701, South Africa; Ian M. Orme, Colorado State University, Fort Collins, CO 80523

Citation: Barbier M, Wirth T. 2016. The evolutionary history, demography, and spread of the *Mycobacterium tuberculosis* complex. *Microbiol Spectrum* 4(4):TBTB2-0008-2016. doi:10.1128/microbiolspec.TBTB2-0008-2016.

Correspondence: Maxime Barbier, maxime.barbier@etu.ephe.fr

that can explain the contrasted infectivity and transmission observed between *Mycobacterium tuberculosis* complex (MTBC) lineages (2–4), without neglecting the genetic architecture of the host's immune system (5).

Another challenge we have to face is the effect of globalization, i.e., the dramatic increase of population and individual movements that encompass touristic activities, refugee diasporas, and, soon to come, climatic migrants. This ongoing maelstrom has multiple consequences, such as an increasing number of patients infected by nonendemic strains, the spread of multidrug-resistant (MDR) strains from health care-deficient countries, and the frightening specter of the expansion of totally drug-resistant (TDR) strains (6). In this review, we will illustrate how genomic insights driven by whole-genome sequencing and comparative genomics can help us to combat this old foe, and unravel its evolutionary history, spread, and demography. From a more practical point of view, the approaches we will discuss here, combined with selection and population genomics models, might also help us to evaluate the impacts of treatment programs on the relative transmission success, to pinpoint the molecular targets of selection, and, eventually, to develop new drugs.

HISTORICAL CONSIDERATION AND EARLY (MIS)CONCEPTIONS ON TUBERCULOSIS EVOLUTION

Few diseases, with the exception of plague (*Yersinia pestis*), have left such an important written signature as tuberculosis. The first literary traces were detected in Chinese medical texts predating the Xia dynasty and in the Indian Vedas (7), respectively, some 5,700 and 3,500 years ago. Until recently, little was known about tuberculosis origins, evolution, and spread. Thanks to the development of molecular tools, four distinct species were identified as causing the disease: *M. tuberculosis*, the human pathogen; *M. bovis* (8), found primarily in cattle; *Mycobacterium africanum* (9), isolated from African patients; and *Mycobacterium microti* (10), isolated from voles. These species were defined based on the host from which strains had been isolated. However, biochemical analyses, including *in vitro* growth rates, microscopic observations, and differential host-specific pathogenicity, suggested that interspecies borders were less well defined than initially expected (11). All these taxa belong to the MTBC, although their status in terms of taxonomic level (species, subspecies) might be further debated. In this group, *M. tuberculosis* and *M. bovis* are the more prevalent ones, although this might be due to strong sampling biases driven by health-economic priorities, as well as by differences in access to funding. *M. tuberculosis sensu stricto* infects humans and, until recently, the species was divided into five variants based on biochemical properties, namely the classical human, Asian human, bovine, African I, and African II variants (12).

The initial paradigm concerning the evolution of *M. tuberculosis* was that the bacillus evolved from *M. bovis* (13). Thanks to novel molecular data, however, this scenario was revised (14), although the old concept keeps being cited (15). The observations that led to these prime conclusions were the following. First, many diseases afflicting humans are zoonoses, and tuberculosis should be no exception. Famous examples of transmission from animals to humans encompass the Ebola virus, HIV, and Chagas' disease (15). The transmission process can oscillate between sporadic outbreaks with little human-to-human transmission and a more settled coevolution if the bug can adapt to its new niche. Based on this knowledge, the initial hypothesis was that a cattle *M. bovis* strain infected a human and successfully spread in the *H. sapiens* populations. After some millennia of coevolution, the bacterium specialized to its novel host, became human specific, and is now known as *M. tuberculosis*. In fact, it is not unusual to see patients infected by bovine tuberculosis, with transmission occurring via aerosols or the consumption of infected milk. Moreover, until now, no human remains older than 11,000 years have shown traces of tuberculosis disease (16, 17), whereas the most ancient animal case has been found in a 17,000-year-old extinct bison (18). This apparent anteriority of animal infection was used to promote the cattle-to-human transmission route hypothesis. Furthermore, the fact that the earliest human remains carrying tuberculosis date back to the Neolithic revolution (8,000 to 10,000 years ago) is intriguing and suggests some causality with the rise of domestication. It is tempting to think that the concomitant increase of animal stocks and host population size favored the interactions and contacts between these two players. Indeed, in the past, humans shared their home with bovines to protect them against predators and extreme temperatures (14): a single infected and coughing animal might have been able to transmit the disease to an entire family.

Here we see the dangerous attraction we have for nice, logical, flowing narratives; yet, mycobacterial interspersed repetitive unit genotyping and whole-genome sequencing (WGS) provided strong evidence against such a linear explanation, as we shall see in the next section.

THE PREGENOMIC ERA AND FIRST-GENERATION PHYLOGENETIC ANALYSES

The Fingerprint Era

The advances of molecular biology enabled the study of bacterial DNA, unraveled fine-scale genetic structures, and clearly segregated sister strains, which previously seemed nearly identical, mostly because morphological and biochemical traits provide little information about relatedness and species phylogenies. We will present first the main pre-next-generation sequencing (NGS) methods that enabled us to discriminate the principal MTBC families and to disentangle their evolutionary link, and how these methods shifted our vision of tuberculosis evolution and spread.

One of the first typing techniques applied to *M. tuberculosis* was the restriction fragment length polymorphism (RFLP) method. It is a fingerprint-based approach that relies on the enzymatic digestion of the circular chromosomal DNA, followed by gel electrophoresis revealed with radiolabeled probes targeting a particular sequence, such as an insertion sequence (IS). The IS*6100* RFLP analysis (19) has been widely used for *M. tuberculosis* molecular typing. The IS*6100* sequence is usually present in multiple copies on the chromosome. Depending on the locations and the copy number of this element, a profile is established that allows for strain discrimination. Such profiles enabled the demonstration that *M. microti* was responsible for human infections (20). Other insertion sequences have been used such as IS*1081* (21) and IS*986* (22), but they did not reach the success of IS*6100*. This marker presents a high evolutionary pace and therefore evolves relatively quickly in an otherwise relatively homogeneous genetic background. Therefore, IS*6100* proved very useful in epidemiological studies and facilitated the segregation of clusters of closely related strains or, in the best case, of clones (23). A major drawback of this technique comes from its poor portability and laboratory dependency in terms of fingerprint profiles, leading therefore to little insight at larger evolutionary scales (24). Besides, differentiation of strains is strongly dependent on the number of IS*6100* copies. Strains with high copy numbers are accurately differentiated from their close variants, while strains with few copies are more difficult to segregate. Numerous other markers have been used in the field of microbiology with more or less the same advantages and flaws (25). Another limitation of RFLP is that it requires mycobacterial culture, lasting from 20 to 40 days. This time frame is very long when studying infection chains in a clinical context. According to a search on the Web of Science database, the topic "IS*6100*" reached a citation apex in 2012 and now follows a gentle but regular decline.

The Multilocus Era

Next came PCR-based techniques that allowed fast, reproducible, and efficient typing (between 1 day and 1 week) like the spacer oligotyping method, called "spoligotyping" (26). The goal of this technique is to type the direct repeat (DR) locus of *M. tuberculosis*. This locus is an alternation of DRs, composed of a well-conserved sequence of 36 bp, and nonrepetitive spacer sequences, 34 to 41 bp long. *M. tuberculosis* strains can be discriminated based on their number of DRs and the presence or absence of particular spacers (27). Spoligotyping is therefore an efficient typing method (28) that differentiates MTBC strains from other environmental mycobacteria and clearly separates *M. bovis* from *M. tuberculosis*. It has less discriminatory power than IS*6100*, when present in high copy numbers, but it is present in all strains, unlike IS*6100*. The principal inconvenience of this method is that spoligo patterns are CRISPR structures that play a role in *Eubacteria* and *Archaebacteria* defense against phages (29). Consequently, this marker is under strong diversifying selection, prone to homoplasy, and of little interest for phylogenetic reconstructions, if any.

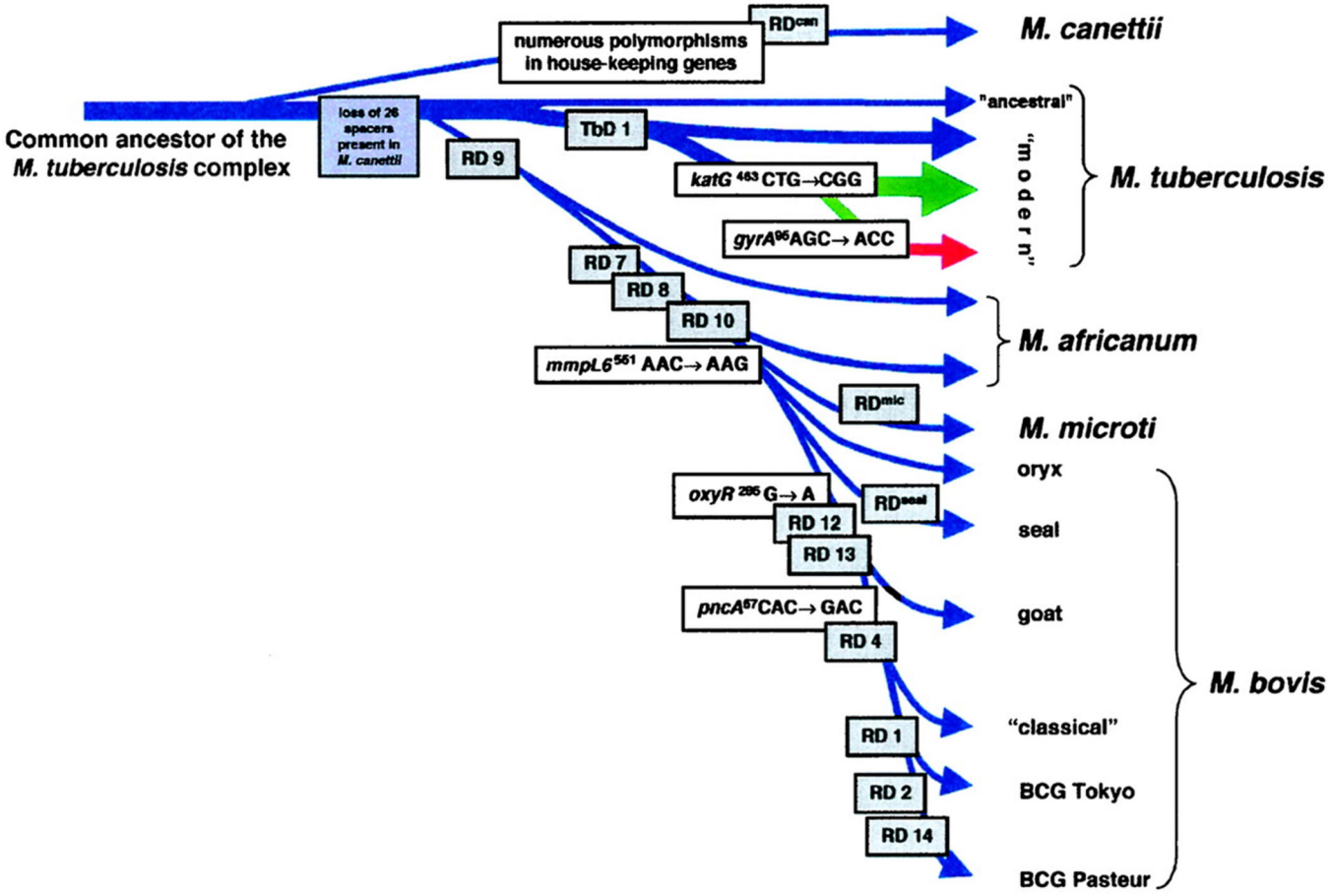
The second PCR-based method developed in the early 2000s is a high-resolution typing method based on variable number tandem repeats (VNTRs) of genetic elements named mycobacterial interspersed repetitive units (MIRUs) (30). Those markers resemble the human minisatellite-like regions developed by Sir Alec Jeffreys and widely used in forensics. MIRU loci are scattered in the genome of *M. tuberculosis* and consist of repetitive patterns 51 to 77 bp long. Only minor indels or polymorphisms occur in these sequences, mostly following the so-called stepwise-mutation model (SMM), meaning that the allelic state changes by the acquisition or the loss of one repetitive unit. The typing of these MIRUs is simply the measure of the number of repetitions at each locus where the number of repetitions varies between 0 and a maximum of 25. The MIRU typing underwent progressive upgrading steps, from 12 to 15 and finally 24 MIRU loci.

A third way used to differentiate strains consisted of the sequencing of a set of structural genes, allowing defining their relatedness based on sequence polymorphisms. Sreevatsan et al. implemented this approach in 1997 (31). They used 26 different genes or gene fragments, in which they observed a lack of neutral mutation, with up to 95% of nonsynonymous mutations associated with antibiotic

resistance. However, phylogenetic inferences and evolutionary scenarios inferred from genes under strong positive selection (involved in antibiotic resistance, for example) are generally not reliable. Alternatively, taking into account the sole synonymous single nucleotide polymorphisms (SNPs) might be a solution but this remains very restrictive. The authors ultimately used the only two nonsynonymous mutations that were not involved in antibiotic resistance to define three different groups. We see here that, because of technical barriers, biased locus choice, and little genetic diversity of the MTBC, the pre-NGS area sequencing studies remained tricky and led to partially misleading conclusions.

Yet another simple analytical approach turned out to be far more promising. Thanks to genome-wide comparisons (32), variable regions resulting from insertion-deletion events have been discovered. Approximately 20 such regions are phylogenetically highly informative since they follow a completely parsimonious and non-homoplastic evolutionary path, from presence to definitive loss, turning any change into a marble-engraved event (33). Based on these so-called regions of difference (RDs), a new evolutionary scenario emerged that contradicted previous thoughts (Fig. 1) because it stated that human strains did not derive from *M. bovis*. RD-based analyses even suggested that humans transmitted

FIGURE 1 Diagram of the proposed evolutionary pathway of the tubercle bacilli illustrating successive losses of DNA in certain lineages (gray boxes). The diagram is based on the presence or absence of conserved deleted regions and on sequence polymorphisms in five selected genes. The distances between certain branches may not correspond to actual phylogenetic differences calculated by other methods. Blue arrows indicate that strains are characterized by *katG*463. CTG (Leu), *gyrA*95 ACC (Thr), typical for group 1 organisms. Green arrows indicate that strains belong to group 2 characterized by *katG*463 CGG (Arg), *gyrA*95 ACC (Thr). The red arrow indicates that strains belong to group 3, characterized by *katG*463 CGG (Arg), *gyrA*95 AGC (Ser), as defined by Sreevatsan et al. (31). Adapted from Brosch et al. (33).



tuberculosis to cattle and other animals rather than the other way around. One deletion, TbD1, separates "modern strains" from "ancient strains." This latter clade comprises animal strains, *M. africanum*, and some less virulent human strains, whereas "modern strains" exclusively infect humans. An interesting analysis of PhoPR virulence factors provided some novel insights into those splits and might explain how some mutations lowered the virulence of strains belonging to *M. bovis* and *M. africanum* (34). As an alternative to overstep the lack of accuracy or the drawbacks of all markers previously presented, some researchers chose to combine them (35). Such combined analyses have been conducted in numerous surveys and allowed the separation of *M. tuberculosis* human strains into different clades (Table 1). Those clades were initially named based on the prevalence and geographical source of their members (36, 37). The most remarkable phylogeographical clades belonging to the "modern strains" were Beijing (highly prevalent in East Asia), CAS (central Asia), and Haarlem, X, and LAM (Latin American-Mediterranean). *M. africanum* and EAI (East African Indian) composed the "ancient strains" group. Captivatingly, lineages from neighboring regions are more closely related than randomly chosen lineages, advocating for a strong biogeographical structuring: Haarlem, X, and LAM clades are more prevalent in Europe and cluster together; the same holds for the Asiatic Beijing and CAS clades. The other lineages form a paraphyletic group of ancient strains, which are essentially restricted to Africa and India. The observed relationships between MTBC lineages are similar to those observed in humans, suggesting that humans could have carried tuberculosis for millennia and that the present-day geographical distribution of tuberculosis has been shaped by ancient if not first human migrations out of Africa. The hypothesis that humans and *M. tuberculosis* coevolved and spread together has been studied and detailed by Wirth et al. (38). Using MIRU genetic markers, the authors identified two major clades, one composed of human strains only and one containing both human and animal strains. Interestingly, the basal and genetically more diverse lineage of the second clade infects humans, confirming that animal strains derived from human ones. Moreover, using Bayesian approaches and coalescent-based theory, they estimated the clade ages and inferred the *M. tuberculosis* demographic history. Based on these calculations, the common ancestor of the MTBC appeared some 40,000 years ago. In a second step, the ancestral strains reached the Fertile Crescent where they diversified during and shortly after the onset of domestication, 10,000 years ago. They ultimately spread out of Mesopotamia, accompanying different human migration waves in Africa, Asia, and Europe, and gave rise to locally adapted pathogens. Furthermore, a strong signal of demographic expansion was detected in the past 200 years, concomitant with industrialization. All these clues point toward a strong association and long coevolution between *H. sapiens* and *M. tuberculosis*.

Last, just before the rise of NGS and WGS, some researchers began to use SNP-based approaches to assess lineage relationships and to unravel deep MTBC sublineages. Since mutations are rare in *M. tuberculosis* genomes, they compared the complete genomes of few available reference strains and identified a list of SNPs. Then they sequenced these genes or called the SNPs in a large NGS data set gathered from strain collections (39–41). The authors retrieved the principal clades described above, but all generated phylogenies turned out to be poorly resolved, ending in star-like topologies. At first glance, one might have invoked a sudden radiative burst and a hard polytomy. What we were facing was, in fact, a methodological issue called ascertainment bias, which is often driven by biased taxonomic sampling (42). Indeed, strong ascertainment bias and related phylogenetic reconstructions systematically lead to the collapse of divergent lineages into single points, failing therefore to generate reliable tree topologies. This is exemplified by the Filliol et al. (41) paper, where the authors reached the unrealistic conclusions that *M. tuberculosis* had an Indian origin.

NGS AND TUBERCULOSIS EVOLUTIONARY HISTORY

The Global Picture

The ultimate knowledge that can be gathered using NGS is a complete list of all nucleotides that constitute the circular chromosome of a strain. Our understanding of the evolutionary relationships of the different MTB lineages, their radiation, and time to the most common ancestor (TMRCA) greatly profited from WGS, resulting in a quantum-leap progress in the field of MTBC phylogenetics (43). NGS favored the characterization of the genetic diversity of an increasing number of strains, covering different lineages and large geographic distributions. Thanks to a high-quality reference genome (44) (Sanger sequenced) gathered from the H37Rv laboratory strain, the scientific community has a template on which Illumina or Roche 454 reads can be mapped. Unraveling the topology of the MTBC tree is also highly dependent on the availability of a reliable outgroup.

TABLE 1 Correspondence table of the MTBC human-adapted strains identified by main typing methods and including the latest nomenclature[a]

Evolutionary age (species)	Lineage name based on LSP/SNP[b]	Lineage and sublineage [RD associated]				Spoligotype family
Ancient lineage (*M. tuberculosis*)	Indo-Oceanic lineage	1 [RD239]	1.1	1.1.1		EAI4 and EAI5
					1.1.1.1	EAI4
				1.1.2		EAI5 and EAI3
				1.1.3		EAI6
			1.2	1.2.1		EAI2
				1.2.2		EAI1
Modern lineages (*M. tuberculosis*)	East-Asian lineage	2	2.1 (non-Beijing)			MANU ancestor and orphan profile
			2.2 (Beijing) [RD105, RD207]	2.2.1 [RD181]		Beijing
					2.2.1.1 [RD150]	Beijing
					2.2.1.2 [RD142]	Beijing
				2.2.2		Beijing
	East African-Indian lineage	3 [RD750]	3.1			CAS except CAS1-Delhi
				3.1.1		CAS1-Kili
				3.1.2	3.1.2.1	CAS2
					3.1.2.2	CAS
	Euro-American lineage	4	4.1	4.1.1 (X-type)	4.1.1.1 [RD183]	X2
					4.1.1.2	X1
					4.1.1.3 [RD193]	X3 and X1
				4.1.2		T1 and H1
					4.1.2.1 (Haarlem) [RD182]	T1 and H1
			4.2	4.2.1 (Ural)		H3 and H4
				4.2.2		LAM7-TUR and T1
					4.2.2.1 (TUR) [RD182]	LAM7-TUR
			4.3 (LAM)	4.3.1		LAM9
				4.3.2		LAM3
					4.3.2.1 [RD761]	LAM3
				4.3.3 [RD115]		LAM9 and T5
				4.3.4 [RD174]	4.3.4.1	LAM1
					4.3.4.2	LAM11-ZWE, LAM9, LAM1, and LAM4
					4.3.4.2.1	LAM11-ZWE
			4.4	4.4.1	4.4.1.1 (S-type)	S
					4.4.1.2	T1
				4.4.2		T1 and T2
			4.5 [RD122]			H3, H4, and T1
			4.6	4.6.1 (Uganda) [RD724]	4.6.1.1	T2-Uganda
					4.6.1.2	T2
				4.6.2 [RD726]	4.6.2.1	T3
					4.6.2.2 (Cameroon)	LAM10-CAM
			4.7			T1 and T5
			4.8 [RD219]			T1, T2, T3, T4 and T5
			4.9 (H37Rv-like)			T1
Ancient lineages (*M. africanum*)	West-Africa lineage 1	5 [RD711]				AFRI_2 and AFRI_3
	West-Africa lineage 2	6 [RD702]				AFRI_1
Intermediary lineage (*M. tuberculosis*)	Lineage 7	7				

[a]Regions of deletion (RD) are given in brackets and appear below the lineage/sublineage in which they are present. Synthetic table adapted from Coll et al. (86).
[b]LSP, large sequence polymorphism.

Fortunately, Supply and colleagues (45) sequenced and analyzed the whole genomes of five strains belonging to the smooth tubercle bacilli (STB), the closest outgroup known so far (46, 47). These strains harbor a unique smooth colony phenotype on culture media, are less persistent and virulent than their *M. tuberculosis* counterpart, and were essentially collected from the Horn of Africa, the cradle of humankind. Furthermore, the so-called "*Mycobacterium canettii*" and/or "*Mycobacterium prototuberculosis*" strains display a unique feature in the MTBC world: they are highly recombinogenic and they are prone to horizontal gene transfer (HGT). Indeed, they possess distinct CRISPR-Cas systems relative to *M. tuberculosis* that are closely related to the genera *Thioalkalivibrio*, *Moorella*, and *Thiorhodovibrio* (45). Interestingly, this latter species is adapted to warm and salty waters, a type of environment that is often encountered in the western part of Djibouti where large saline lakes coexist with hot springs. These scars of past genetic exchanges definitively advocate for an environmental origin of the ancestor of *M. tuberculosis* that might date back 3 million years (46).

Once rooted with *M. canettii*, the first attempt to solve the evolutionary history of the MTBC with full genomes relied on a set of 25 *M. tuberculosis* strains representing the six main human lineages known at that time (48). The molecular diversity of those strains remained rather modest, with only one SNP call for every 3 kb of sequence generated, highlighting the relative youth of this human pathogen, its clonality, and putative rise through a major bottleneck. The neighbor-joining tree proposed by the authors did not add much in terms of branching order but illustrated the power of genomics in terms of bootstrap branch supports (≥99%) and within-lineage resolution. Three major clades could be distinguished, one encompassing lineages 2, 3, and 4; followed by its sister group, lineage 1; and, finally, a marginally more basal group represented by the two *M. africanum* lineages (Table 1). Because the rationale of this first genomic paper was to study *M. tuberculosis* human T-cell epitopes, the absence of animal strains was not surprising. However, in terms of evolutionary history, this no-attendance needed to be corrected in future studies. This was done in a landmark study (49) where genomes of 259 *M. tuberculosis* strains were analyzed. At such scales, with more than 30 strains per major lineage, the likelihood to get much closer to the real picture significantly increases. Comas and colleagues included in this study a new member of the MTBC, the recently described lineage 7 (50), which was only collected from Ethiopian patients, as well as a couple of animal strains. Again, the maximum-likelihood tree confirmed the monophyly of the modern strains (L2, 3, and 4), but also suggested that the animal lineage diverged from the African lineage 6 (*M. africanum*). Another important feature is the strong biogeographical structure of the different lineages; their distribution around the planet is not random at all, and clearly corresponds to well-defined geographic and cultural areas. This observation, coupled with the fact that tree topologies and geographic distribution between MTBC strains and the main human mitochondrial macrohaplogroups were highly similar, prompted the authors to calibrate the tuberculosis evolutionary tree on its human backbone. More specifically, the striking resemblance and branching order of the Southeast Asian and Oceania tuberculosis strains and the Southeast Asian, Oceanian macrohaplogroup M in humans were used for this purpose (Fig. 2).

This elegant approach, coupled with a coalescent-based approach, indicates that the MTBC emerged at least 70,000 years ago. The demographic success and the timing of the propagation of *M. tuberculosis* were evaluated with Bayesian skyline plots (51–53) that unraveled the effective population size of the bug through time. According to this scenario, MTBC accompanied the migrations of anatomically modern humans out of Africa and started to spread at a higher pace during the Neolithic demographic transition (54). It is tempting to connect a sustainable infectious cycle with the advance of farming and domestication, accompanied by dramatic changes in lifestyle, from hunter-gatherers to farmers, from low-density populations to local crowds. However, the data also show that the conquest of the Indian Ocean areas by lineage 1 largely predated the Fertile Crescent onset, starting as early as 67,000 years ago and followed by a second wave of peopling that reached the Middle East, Europe, and Asia some 46,000 years ago. Overall, WGS highlights the coevolution between a host and its pathogen, *H. sapiens* and *M. tuberculosis*, their intricate evolutionary histories, their African origin, and their adaptation from low to high population densities.

Yet recently a new publication dramatically affected the temporal dimension of the scenario presented above. Bos et al. (55) analyzed three 1,000-year-old mycobacterial genomes from Peruvian skeletons showing stigmata of tuberculosis infection that proved to be *Mycobacterium pinnipedii* (Fig. 3), a type of strain mostly isolated from seal species in the Southern Hemisphere. It is worth mentioning that two of the archeological sites (El Algodonal and Chiribaya Alta) were close to the Rio

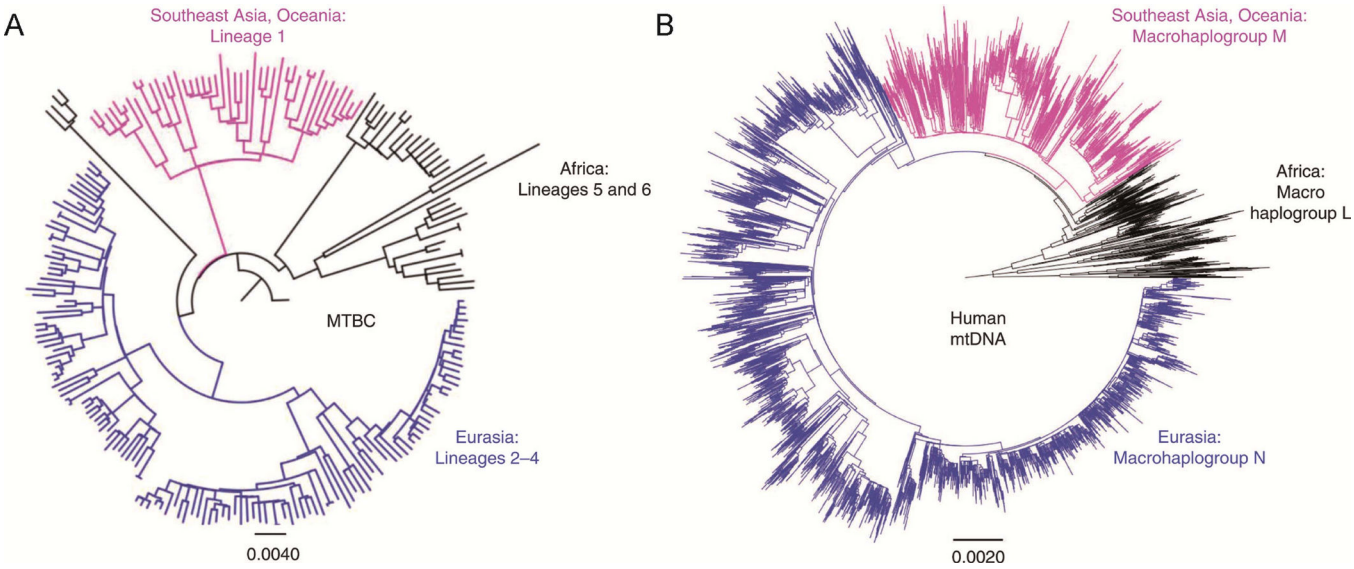


FIGURE 2 The genome-based phylogeny of MTBC mirrors that of human mitochondrial genomes. Comparison of the MTBC phylogeny **(A)** and a phylogeny derived from 4,955 mitochondrial genomes (mtDNA) representative of the main human haplogroups **(B)**. Color-coding highlights the similarities in tree topology and geographic distribution between MTBC strains and the main human mitochondrial macrohaplogroups (black, African clades: MTBC lineages 5 and 6, human mitochondrial macrohaplogroups L0 to L3; pink, Southeast Asian and Oceanian clades: MTBC lineage 1, human mitochondrial macrohaplogroup M; blue, Eurasian clades: MTBC lineage 2 to 4, human mitochondrial macrohaplogroup N). Scale bars indicate substitutions per site. Adapted from Comas et al. (49).

Algodonal and only 5 to 10 km upstream from the river mouth. The team led by German experts in ancient DNA managed to successfully sequence these genomes by applying DNA capture (56) and genomic assembly of the metagenomic reads. The assembled genomes harbored the typical signature and damage of ancient DNA, and accounted for 2% of the total reads. The authors then calibrated the molecular clock using the archeological data and the fact that branch lengths are a function of the elapsed time, being longer for strains collected in the 21st century and shorter for much older strains. This Bayesian calibration process, under a relaxed clock model, resulted in a substitution rate of about 5×10^{-8} substitutions per site per year, placing the most recent common ancestor for the MTBC at 4,000 years, which turns to be more than one order of magnitude younger than the age proposed by Comas and colleagues (49). For comparative purposes, we should mention that the substitution rate obtained by Comas et al. was much slower, i.e., 2.6×10^{-9} substitutions per site per year. This MTBC TMRCA dating issue definitively splits the mycobacteriology community into two entities, i.e., the pros and the cons. The more recent dating of the Bos study conflicts with numerous archeological proofs, including evidence of MTBC in a 17,500-year-old bison in Wyoming, United States (18), the presence of a 9,000-year-old modern tuberculosis strain in a Neolithic infant skeleton from Israel (16), and an animal MTBC strain harboring the RD9 deletion some 7,000 years ago (57). The cumulative evidence gathered from amplification of IS*6100* and spoligotyping patterns from bones predating the Bos et al. MTBC TMRCA are questioned by some scientists, claiming that these mobile genetic elements and CRISPR systems are not MTBC specific enough and that they might be observed in environmental mycobacteria (58, 59), leading to false positives. The same arguments are used to question the validity of the presence of mycolic acids (60, 61) in biosamples to identify MTBC strains. Other colleagues came to the same molecular clock as Bos et al. (55) using a calibration point based on aboriginal communities in Canada that acquired *M. tuberculosis* via the fur trade in the late 18th century (62). Nearly identical rates were obtained again based on a calibration relying on an 18th-century mummy collected from a Dominican church in Hungary (63). Furthermore, Pepperell et al. (62) did not find statistical support for codivergence of *M. tuberculosis* with its host in formal phylogenetic congruence tests.

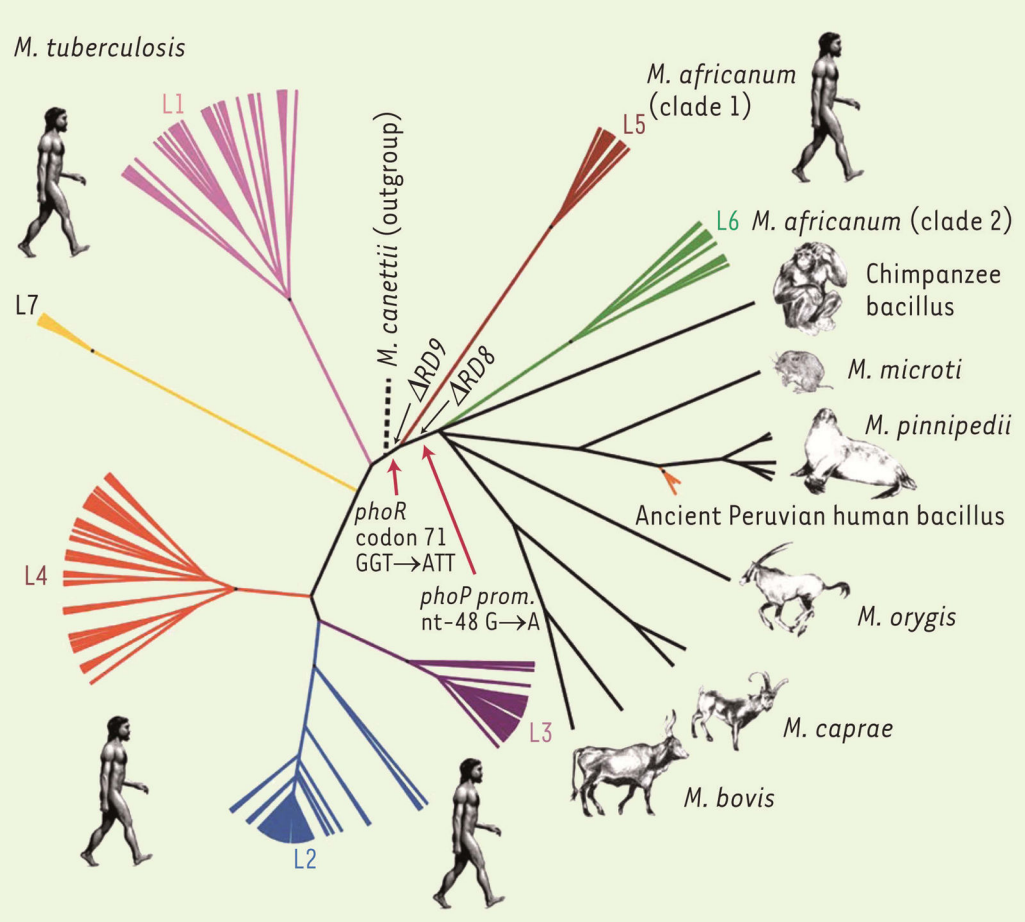


FIGURE 3 Whole-genome phylogeny of 261 strains belonging to the MTBC. Animal and *M. africanum* specific deletions are indicated, as well as mutations affecting the PhoPR virulence regulator. Adapted from Bos et al. (55) and Gonzalo-Asensio et al. (34).

Another by-product of this study is that, according to the authors, seals are the source of New World human tuberculosis, therefore predating the likely entry of tuberculosis in South America with the Conquistadores, putatively harboring lineage 4 strains in their lungs. The Bos et al. conclusions based on the sole observation of a couple of ancient Peruvian humans are overinterpreted in the best case if not dubious at all. This scenario possibly transforms the exception into the rule by extrapolating conclusions based on local observations to continental-scale lessons. A more parsimonious setup could be built on a small population of indigenous hunters who incidentally contracted tuberculosis from infected seals that might be part of their natural prey or diet. Such transfers are rare but can be observed in zoos; notably, South American sea lions managed to infect a camel and a Malayan tapir in neighboring enclosures with their *M. pinnipedii* strains (64), and animal keepers were infected in a zoo in the Netherlands (65). The observation made at the southern border of Peru might be therefore anecdotal and may have resulted in an evolutionary dead end. Before claiming that *M. pinnipedii* plagued South and North American indigenous populations, before being completely replaced by the L4 lineage in present days, far more evidence is needed. This includes additional samples from a larger geographic distribution and additional workable skeletons from diverse archeological sites.

Animal-Related MTBC Strains

According to the currently available sampling and population genomics data, the animal lineages emerged from a common ancestor closely related to lineage 6 (*M. africanum*) (55) (Fig. 3). Consequently, multiple mammalian host jumps occurred leading to adaptive processes and genomic erosion (66). Those animal genomes are of particular interest because genes that undergo pseudogenization or get lost are indicative of

host specificity, notably here *H. sapiens* specificity. Interestingly, three independent losses of the RD1 region have been observed in *M. microti* (67), the dassie bacillus, (68) and *Mycobacterium mungi* (69). This convergent evolution might underline the key role of the ESX-1 secretion system for infecting *H. sapiens*. Beyond the evolutionary dimension, the adaptive radiation of animal MTBC should attract more attention since animal lineages can tell us a lot about human-specific genes, which are prone to be altered or deleted in genomes belonging to the former lineages. This critical situation is illustrated by the relative paucity of published animal MTBC strains, with the notable exception of *M. bovis*, where veterinary and socioeconomic factors prevail. The few available genomes cover the following members of MTBC, *M. suricattae*, the chimpanzee bacillus, *M. microti*, *M. pinnipedii*, *M. bovis*, and *M. caprae*, but no phylogeny including all these members has been published so far. The likelihood that other ignored animal lineages exist in the field is high; a good hint would be to further investigate in the direction of social or highly promiscuous mammal species where the settlement of epidemic episodes are favored.

Zooming into the Lineages

One of the major advances linked with WGS is the possibility to switch to population genetic approaches in the field of *M. tuberculosis* since enough SNPs can be accumulated in the evolutionary history and the coalescence of local populations. For instance, up to 0.4 mutations per genome per year can be accumulated (70). Applying such an approach, Comas et al. (49) scrutinized the evolutionary history of the Beijing lineage, an important member of lineage 2. The Beijing family attracted much attention because its members are hypervirulent in mouse models, spread quickly in Eurasia and Western Europe, and are associated with multidrug resistance (71). The family TMRCA was estimated at 8,000 years coinciding with the rise of agriculture in the Yangtze River region, the domestication of crops and the onset of Chinese farmer populations. Interestingly, the Bayesian skyline of the Beijing family matched pretty well the one obtained from the human mitochondrial haplogroups from East Asia, confirming this likely scenario. Moreover, the dating of the Beijing family is relatively congruent with former analyses based on MIRU typing (38), but also with more recent data obtained from a large collection of 5,000 Beijing strains (72). In the later publication, Merker et al. (72) unraveled the genetic structure and global spread of the Beijing lineage; this lineage is globally distributed but still entails a fine-scale genetic structuring. The authors detected six clonal complexes (CCs) and one basal lineage; those CCs proved to be strongly associated with geographical entities (Fig. 4). They also confirmed that this lineage initially originated in the Far East 6,600 years ago from where it radiated worldwide in several waves. This was illustrated by a negative correlation (r^2 = 0.626) between the mean allelic richness of the strains and their distance from the Yangtze River. An ancestral East Asian population of strains, mostly endemic, that gave rise to new variants following different migration routes, can explain this pattern. The consequence of this scenario is a stepping-stone propagation of the germs, followed by successive bottlenecks, resulting in genetic erosion with increasing distance from the source. The situation is similar for *H. sapiens* and its little companion *Helicobacter pylori*, where the highest genetic diversity can be observed in Africa and the lowest one in South America (73, 74). Worth mentioning are the contrasted profiles between the ancestral strains (CC6 and BL7) that only marginally dispersed from their area of endemicity and the other derived CCs that successfully spread at continental scales. CC5 is probably the best illustration to show how a minor variant, originating from Southeast Asia, spread some 1,500 years ago into the Pacific and increased its frequency due to drift and successive founder effects, culminating at a more than 90% prevalence in Micronesia and Polynesia. One of the most striking features was the evolutionary history of CC1, also called the central Asian clade, and CC2, the Russian clone. The first CC spread westward, becoming highly predominant in central Asia and around the Black Sea, and the latter one became predominant in Russia and Eastern Europe. Both CCs had the highest clustering rates for MDR strains, indicating population expansion amplified by the recent transmission of MDR strains. The demographic success of CC1 and CC2 was confirmed by coalescent-based analyses, and their expansion dated back some 200 to 250 years ago.

These recent expansions remarkably match known episodes of migration in Asia. Indeed, several waves of Chinese refugees migrated to the Russian empire, especially Kyrgyzstan, Kazakhstan, and Uzbekistan, as a consequence of a series of national uprisings from 1861 to 1877, which might have driven the expansion of the CC1 and CC2 strains in these regions (75). These recent western expansions are probably superimposed on a more historical, continuous flux of the different Beijing sublineages westward along the Silk Road. After a tip-dating calibration, the authors reconstructed the demogenetical changes in the Beijing lineage based on a

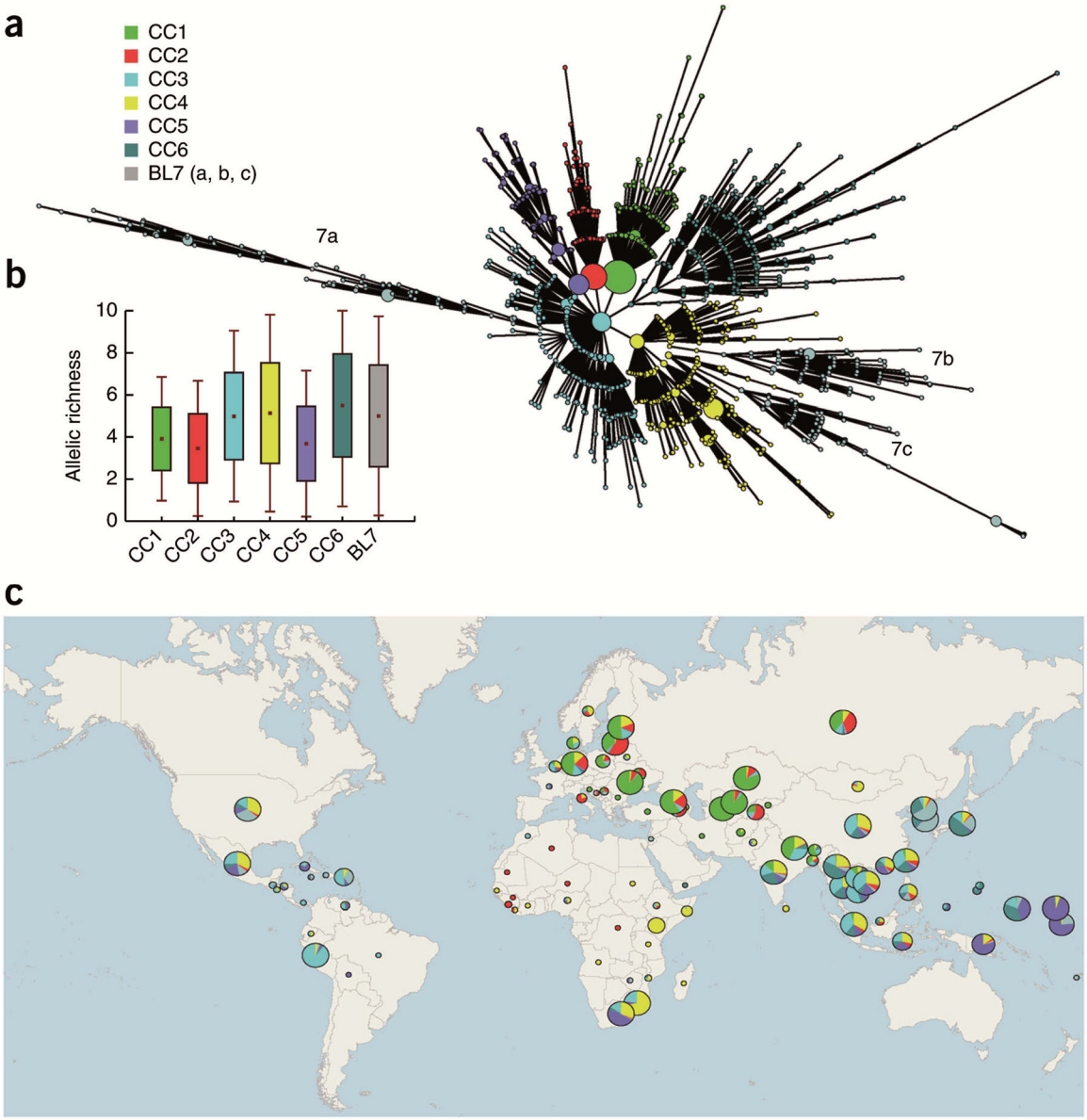


FIGURE 4 Biogeographical structure of the *M. tuberculosis* Beijing lineage. **(a)** MStree based on 24 MIRU-VNTR markers delineating the clonal complexes (CCs) gathered from a worldwide collection (n = 4,987). Major nodes and associated multilocus variants were grouped into six CCs and a basal sublineage (BL). **(b)** Genetic variability in the different Beijing lineage CCs and the BL calculated using a rarefaction procedure. Dots correspond to the mean allelic richness; boxes correspond to mean values ± standard error of the mean and error bars correspond to mean values ± standard deviation. **(c)** World-wide distribution of the Beijing CCs and BL. Each circle corresponds to a country, and circle sizes are proportional to the number of strains. Adapted from Merker et al. (72).

subset of 110 genomes, and they detected a two-step increase in *M. tuberculosis* population size. The first expansion corresponded with the industrial revolution and the second one took place at the end of the 19th century, fitting the information gathered from historical and medical records. This trend was more pronounced for the most-westward distributed clonal complexes and the combined epidemic growth periods resulted in a

10-fold increase of *M. tuberculosis* Beijing's effective population size. The mild population drop that followed the expansion phase took place in the early 1960s and might be linked to the democratization of the antibiotic use. The analysis also captured a last tiny population growth that matches the rise of the HIV epidemics. To conclude, this study demonstrated the power of NGS to explain past and yet "uncaptured" migratory paths from a single lineage and how societal changes impact tuberculosis demography and epidemics. The sudden success of some lineages can even result in full lineage replacements, as exemplified in other pathogens such as *Salmonella enterica* serovar Typhi (76), highlighting the need for and power of population genomics. In the same year, Luo et al. (77) analyzed a novel data set comprising whole-genome sequences of 358 East Asian strains belonging to the Beijing family. The authors applied the molecular clock of Comas et al. (49), which is much faster than the mutation rate implemented by Merker et al. (72). Consequently, both scientific teams reported a demographic expansion of the Beijing family, similar genetic structure and spread, but strongly disagreed on the timing and TMRCA calculations. This situation might be disturbing for the nonspecialist, and it deserves specific explanations. The molecular clock issue will be addressed in greater detail in "The relativity of the clock" (see below), which might help to clarify the situation and propose analytical improvements.

Another lineage that attracted much attention is lineage 4 (the Euro-American lineage) that circulates in Aboriginal and French Canadian communities (78). Some sublineages were introduced in the indigenous populations in the mid-18th century and spread westward through canoe routes until 1850, illustrating the impact of recent trans-Atlantic migrations on remote North American communities. It is particularly worrying to see that the Inuit living in the Nunavik region of Québec present an incidence 50-fold higher than the Canadian average. In a recent population genomics analysis, Lee et al. (79) disentangled the genetic diversity and population structure of 163 *M. tuberculosis* strains scattered in 11 remote Inuit villages. Their main finding confirmed that all patients harbored either one of two sublineages belonging to lineage 4; the TMRCA of the main sublineage, represented by 94% of the strains, dated back to the early 20th century. This result shows that the spread of tuberculosis was not interrupted after the fur trade decline, but that indigenous communities are still prone to "foreign"-mediated epidemics.

If we focus at microevolutionary scales, we reach the borders of the molecular epidemiology field and identification of transmission chains (80, 81). Lee et al. (79) nicely showed that pairs of isolates within villages had significantly fewer SNPs than pairs from different villages (6 versus 47), hinting toward intravillage chain transmissions.

Toward a Universal Taxonomic Nomenclature?

One of the major difficulties that a nonspecialist faces when he or she goes through the tuberculosis literature is the fluctuating and evolving nomenclature concerning the different lineages (see Table 1). The nomenclature was mainly driven by a couple of leading teams, starting from phage typing (82), regions of difference parsimony analyses (3), MIRU cladograms (35, 38), extended MLST trees (83), SNP sets (41, 84, 85), and ultimately whole-genome-based phylogenies. With the drop of the costs of Illumina and PacBio sequencing, thousands of new genomes became available, leading to the discovery of fine-scale phylogenetic structuring but also to the unearthing of new lineages. To facilitate the navigation in this growing complexity, Coll et al. (86) proposed a novel SNP-based bar code approach and implemented the PhyTB tool related to the PhyloTrack library. This numeric code relies on a subset of 62 canonical SNPs gathered from essential genes under negative selection that resolves all seven lineages and another 55 sublineages (Fig. 5). This approach can be upgraded and can evolve with the ongoing sequencing effort.

THE RELATIVITY OF THE CLOCK

Substitution Rate Estimates

Deciphering the evolutionary history of tuberculosis is highly dependent on a rigorous estimation of the molecular clock. One effective way to estimate the substitution rate is to focus on recent epidemics linked to a clone (70), retrospective observational studies (80, 87, 88), or even better on measuring the pace of mutational events within a host (89). The concept behind such approaches is that *M. tuberculosis* is composed of "measurably evolving populations" (90–92), meaning that whole genomes accumulate novel mutations over time frames of months to years. Convincingly, all these WGS studies reported congruent estimates of 0.3 to 0.5 SNP per genome per year, which translates roughly into 1×10^{-7} substitutions per nucleotide per year, with no notable difference between hosts (human or macaque). This mutation rate places *M. tuberculosis* at the lower bound of bacterial species, compared with *Staphylococcus aureus* displaying a mutation rate of 1 to 2×10^{-6}

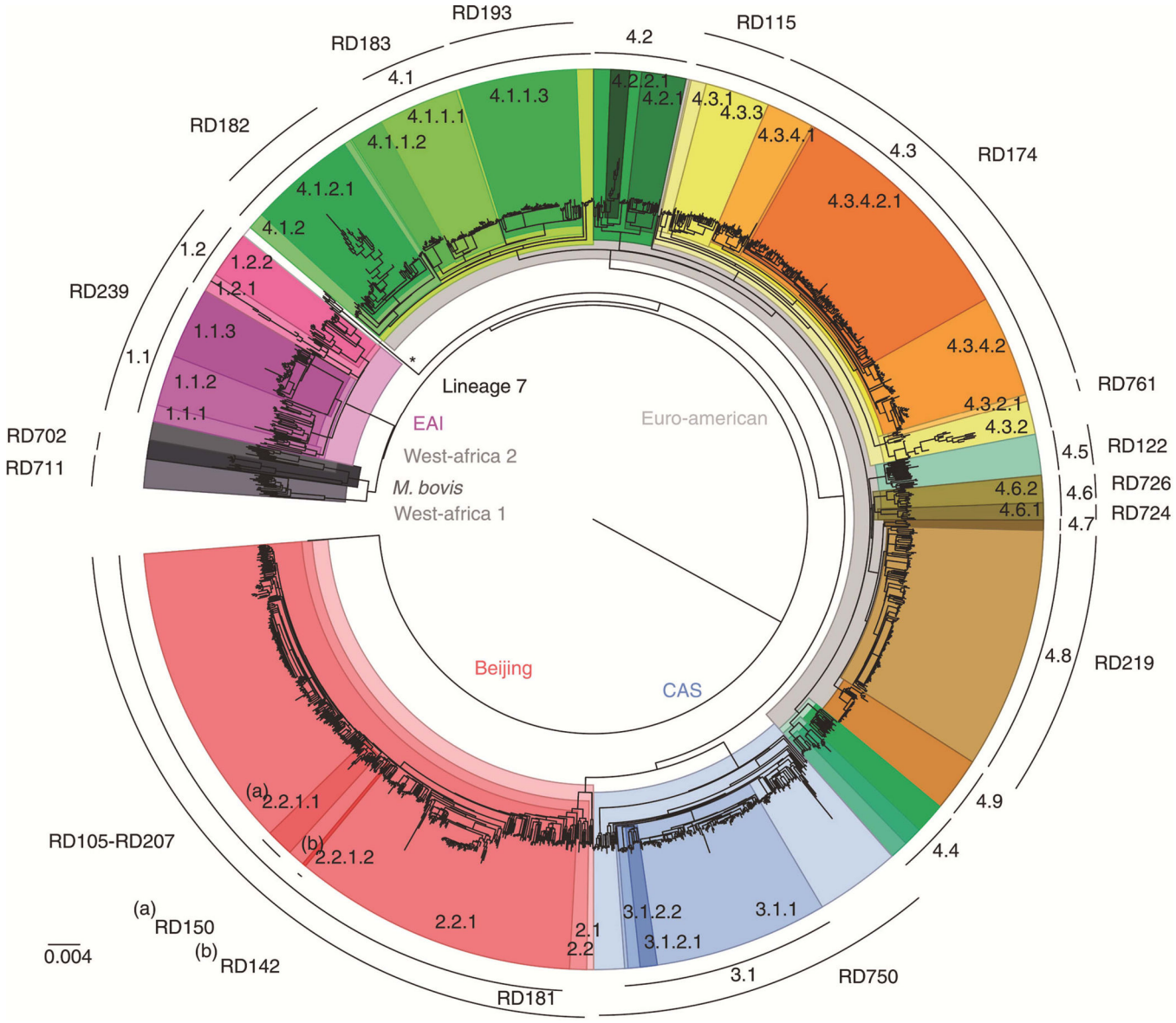


FIGURE 5 Global phylogeny of 1,601 MTBC isolates inferred from a total of 91,648 SNPs spanning the whole genome. All seven main MTBC lineages are indicated in the inner area of the tree. The main sublineages are annotated at the outer arc along with lineage-specific RDs. Identified clades are color-coded. Adapted from Coll et al. (86).

(93, 94), *Escherichia coli* of 5×10^{-6} (95, 96), and the mismatch repair-lacking *H. pylori* of 1×10^{-5} (97, 98) to 7×10^{-4} substitution per nucleotide per year during the acute phase of infection (99). Another approach to calibrate the clock relies on the high similarity of the human mtDNA-based phylogenies and the MTBC human-specific phylogeny, anchoring the Southeast Asian Oceanian lineage 1 with the human macrohaplogroup M (49). This alternative strategy resulted in a substitution rate estimate of 2.58×10^{-9} substitutions per site per year. These two substitution rates are rather incompatible and divergent. So the question is, how can they be combined into a single model?

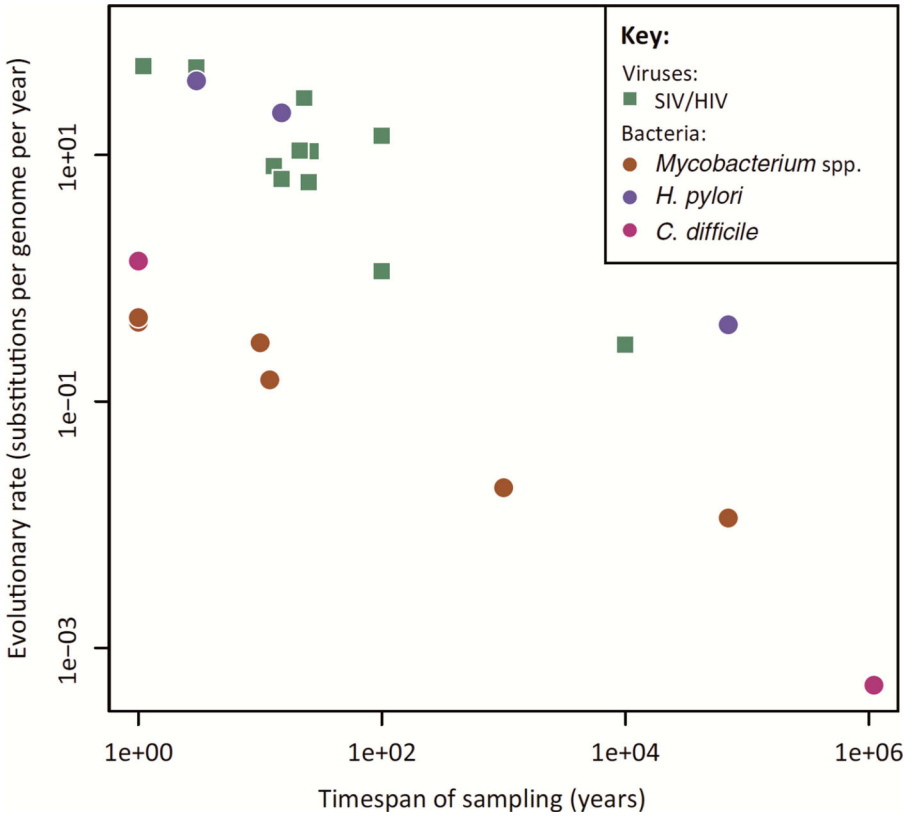
A way to present the problem is to invoke the fields of quantum physics and relativity to build a couple of metaphors. For example, the observer effect and the Heisenberg uncertainty principle stipulate that there is trade-off in capturing simultaneously the position and the momentum of a particle, meaning that obtaining the exact position will lower the information concerning the momentum. In the same way, applying a short-term mutation rate to a *M. tuberculosis* data set covering

a large temporal scale will likely provide reliable information concerning the terminal nodes and the demogenetic changes in the past century, but will perform poorly in terms of TMRCA and vice versa. The other metaphor concerns the theory of relativity where the faster the relative velocity is, the greater the magnitude of time dilatation will be. Again, here we can imagine that the substitution rate is a function of time and may vary following a yet-to-discover mathematical law.

These concepts have indeed some biological meaning, as we shall see. The substitution rate refers to the rate at which nucleotide changes become fixed in populations. This notion differs from the mutation rate, i.e., the rate at which novel mutations arise. In the latter case, some slightly deleterious mutations will be progressively removed by purifying selection, gradually in large populations and more stochastically in small ones. Accordingly, the mutation rate corresponds to the upper limit of mutational changes acquired per unit of time in a given biological system (100). Therefore, the combination of selective pressure (purifying selection), possible saturation at variable sites, and demographic fluctuations will shape the time dependency of evolutionary rates. This trend was noticed based on strong discrepancies between molecular and paleontological dating (101), reviewed by Ho et al. (102), but also subject to some controversies (103). However, there is growing empirical evidence for an exponential-decay law of the substitution rate, fluctuating between two natural boundaries, the mutation and the long-term substitution rates, as exemplified in New Zealand fish species (104), birds and primates (105, 106), and *Vibrio cholerae* (107). This pattern might also be more effective and important in the relative short term (over centuries) for bacterial species and viruses, since they possess much shorter generation times than, e.g., large vertebrates. This is exemplified in Fig. 6, where a strong negative linear correlation between the evolutionary rate and time (both log-transformed) could be detected in three bacterial species, based on available complete genomes.

Consequently, there is an urgent need to reunify the analyses obtained from different scientific teams

FIGURE 6 Consistent with a general pattern for measurably evolving populations, the evolutionary rates of microbial pathogens decrease as a function of the time span over which they are estimated. Data shown are selected representative examples, including one group of RNA viruses and several bacterial pathogens. Adapted from Biek et al. (100).



implementing contrasting short- or long-term substitution rates on *M. tuberculosis* data sets (108) under a same mathematical model. To reach this goal, we will have to define the parameters that define the J-shaped curve of the etiological agent of tuberculosis and to develop, and extend, tools like BEAST that will integrate in a unique coalescent framework, a substitution rate dependent on time (109). However, this task remains challenging, since there is a paucity of reliable calibration points for the intermediate time frames.

Other Limitations

Some additional features must be mentioned that possibly complicate the clock estimates. Among them, we have to consider both intrachromosomal and interlineage variations in the ticking rates. The 4.4 Mb *M. tuberculosis* genome presents highly variable repetitive genetic regions, encompassing genes such as the PPE, PE_PGRS, and ESX families. Those gene families are prone to increased mutation rates, are difficult to assemble, and are often removed from the analyses. Therefore, the accuracy of the trimming step might explain some outliers observed in terms of mutation rate estimates. Furthermore, Martincorena et al. (110) detected mutational hot and cold spots across 2,659 genes from a collection of 34 *E. coli* strains. Lower rates were observed in highly expressed genes and there is no strong argument that *M. tuberculosis* should behave differently. At shorter timescales, mutations affecting genes involved in MTBC DNA repair (111) can inflate the mutation rate, resulting in hypermutator phenotypes. These transient phenotypes are adaptively cardinal and are involved in the fast acquisition of SNPs conferring resistance to second-generation antibiotics. For, example, one copy of the major replicative DNA polymerase, dnaE2, proved to be a mediator of *M. tuberculosis* survival through inducible mutagenesis and contributed to the emergence of drug resistance *in vivo* (112). At a higher hierarchical level, there is growing evidence that some Beijing sublineages undergo faster evolutionary rates relative to other human tuberculosis lineages. Ebrahimi-Rad et al. (113) detected alterations in *mut* genes, mostly missense mutations that improve the adaptability of the W-Beijing clade. In the same vein, Ford et al. (114) demonstrated that cultured *M. tuberculosis* strains from lineage 2 acquired *in vitro* drug resistance against isoniazid and ethambutol three times faster than *M. tuberculosis* strains from lineage 4, invoking again contrasted mutation rates between lineages. However, this difference remains relatively modest compared with *E. coli* (115) and *Pseudomonas aeruginosa* (116), where mutator phenotypes are orders of magnitude more mutable than the wild-type strains. The latter species harbor a mismatch-repair system whose dysfunction increases mutation and recombination rates; in that respect, they differ from *M. tuberculosis*, which is lacking such a system (117).

A final complication can be called up, linked to the peculiar strategy employed by the tubercle bacillus to survive in the host. During latent infection, which accounts for most of its life history, *M. tuberculosis* is in a dormant state with little to no replication activity. This latent stage dramatically contrasts with the active stage of the disease in which clonal multiplication occurs in the lung. The question can therefore be raised whether mutation rates differ between the two different stages and whether this might impact the evolutionary history of the MTBC and the TMRCA estimates. This problem was targeted by a team at the Harvard School of Public Health (89); the authors were able to show, based on whole-genome sequencing of *M. tuberculosis* isolated from cynomolgus macaques, that the mutation rates were similar during latent and active disease. Since most of the chromosome mutations appear during the replication process, this result was somehow unexpected. One of the explanations proposed relies on the high oxidative stress that the bacilli undergo in the macrophage phagolysosome resulting in cytosine deamination or the formation of 8-oxoguanine (118, 119). Alternatively, *M. tuberculosis* keeps dividing more actively during latency than previously thought.

PERSPECTIVES

High financial investments coupled with novel methodological and analytical approaches have driven human population genomics to the upper edge of scientific excellence. The study of the natural host of *M. tuberculosis* reached a new summit with the rise of the study of ancient DNA (120). In 2006, Svante Pääbo and his colleagues analyzed 1 million bp of a 38,000-year-old Neanderthal fossil using the 454 technology (121) and were able to handle the small degraded DNA molecules. Among the most common "lesions," depurination and deaminated cytosine residues near the fragment ends were reported (122). Four years later, thanks to a methodological switch toward Illumina sequencing, they provided the scientific community with the first draft genome of a Neanderthal gathered from three individuals that lived in the Vindija cave in Croatia (123). This major technological step, accompanied with suites of bioinformatics tools, opened the door to late Pleistocene

genomics (124–126) and allowed the discovery of new Hominin lineages, like the Denisovan (127), an extinct relative of Neanderthals.

The first bacterial paleogenomes came just after, notably with the first draft genome of *Y. pestis*, the etiological agent of the plague (128). This ancient genome at an average 30-fold coverage was isolated using targeted capture from Black Death victims in London who died in approximately 1348 to 1350. Even more impressive was the publication of a Bronze Age *Y. pestis* genome from Asia, 3 millennia earlier than any historical records of plague (129). Auspiciously, the bacteria belonging to the MTBC present a complex waxy and hydrophobic cell wall, facilitating therefore the conservation of the DNA molecules. With the development of the field of paleomicrobiology we might soon have the unique opportunity to sample temporal series of skeletons showing stigmata of tuberculosis infection (up to 5% of the patients). Thanks to precise radiocarbon dating, we know that such a material extends through the supposed "life span" of the disease, from the late Pleistocene to the present (17, 130–133). The accumulation of heterochronous *M. tuberculosis* paleopopulation genomics will ultimately solve the molecular clock issue and possibly reunify the fields of evolutionary mycobacteriology, paleoepidemiology, and paleopathology. The first steps have been already accomplished, with 100- (56), 250- (63), and 1,000-year-old (55) genomes made available.

Beyond this rather technical issue, the spread of different lineages through time, lineage replacements, extinct lineages discovery, and eventually a complete revision of current models will be driven by paleomicrobiology. Once Pandora's box is open, we might consider the possibility that Neanderthals may have faced *M. tuberculosis*. If true, two conflicting scenarios can be tested. Neanderthals harbored their own MTBC-like ancestor or they contracted a *H. sapiens*-associated strain during their coexistence with modern humans, some 30,000 years ago in southwestern Europe. Last but not least, ancestral genomes will also improve our understanding of pathogen adaptation and the pace and dynamics of coevolution with its natural host. Studying tuberculosis remains a challenging task, but for sure, we live in interesting times.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions and comments of Jean-Philippe Rasigade, Olivier Dutour, and Jan Willem Dogger on the manuscript. We also thank the Ecole Pratiques des Hautes Etudes for financial support via the priority research action Grant (ARP).

REFERENCES

1. Wirth T, Meyer A, Achtman M. 2005. Deciphering host migrations and origins by means of their microbes. *Mol Ecol* **14:**3289–3306 http://dx.doi.org/10.1111/j.1365-294X.2005.02687.x.

2. Albanna AS, Reed MB, Kotar KV, Fallow A, McIntosh FA, Behr MA, Menzies D. 2011. Reduced transmissibility of East African Indian strains of *Mycobacterium tuberculosis*. *PLoS One* **6:**e25075 http://dx.doi.org/10.1371/journal.pone.0025075.

3. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* **103:**2869–2873 http://dx.doi.org/10.1073/pnas.0511240103.

4. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, Domenech P, Zwerling A, Thibert L, Menzies D, Schwartzman K, Behr MA. 2009. Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J Clin Microbiol* **47:**1119–1128 http://dx.doi.org/10.1128/JCM.02142-08.

5. Brites D, Gagneux S. 2015. Co-evolution of *Mycobacterium tuberculosis* and *Homosapiens*. *Immunol Rev* **264:**6–24 http://dx.doi.org/10.1111/imr.12264.

6. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, Ziazarifi AH, Hoffner SE. 2009. Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. *Chest* **136:**420–425 http://dx.doi.org/10.1378/chest.08-2427.

7. Prasad PV. 2002. General medicine in Atharvaveda with special reference to Yaksma (consumption/tuberculosis). *Bull Indian Inst Hist Med Hyderabad* **32:**1–14.

8. Karlson AG, Lessel EW. 1970. *Mycobacterium bovis* nom. nov. *Int J Syst Evol Microbiol* **20:**273–282.

9. Castets M, Boisvert H, Grumbach F, Brunel M, Rist N. 1968. [Tuberculosis bacilli of the African type: preliminary note]. *Rev Tuberc Pneumol (Paris)* **32:**179–184.

10. Reed GB. 1957. *Mycobacterium microti*, p 703. *In* Breed RS, Murray EGD, Smith NR (ed), *Bergey's Manual of Determinative Bacteriology*, 7th ed. The Williams and Wilkins Co, Baltimore.

11. Tsukamura M, Mizuno S, Toyama H. 1985. Taxonomic studies on the *Mycobacterium tuberculosis* series. *Microbiol Immunol* **29:**285–299 http://dx.doi.org/10.1111/j.1348-0421.1985.tb00827.x.

12. Collins CH, Yates MD, Grange JM. 1982. Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. *J Hyg (Lond)* **89:**235–242 http://dx.doi.org/10.1017/S0022172400070765.

13. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. 2009. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* **7:**537–544 http://dx.doi.org/10.1038/nrmicro2165.

14. Stead WW. 1997. The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin Chest Med* **18:**65–77 http://dx.doi.org/10.1016/S0272-5231(05)70356-7.

15. Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature* **447:**279–283 http://dx.doi.org/10.1038/nature05775.

16. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY, Gernaey AM, Galili E, Eshed V, Greenblatt CL, Lemma E, Bar-Gal GK, Spigelman M. 2008. Detection and molecular characterization of 9,000-year-old *Mycobacterium tuberculosis* from a Neolithic settlement in the Eastern Mediterranean. *PLoS One* **3:**e3426 http://dx.doi.org/10.1371/journal.pone.0003426.

17. Baker O, Lee OY, Wu HH, Besra GS, Minnikin DE, Llewellyn G, Williams CM, Maixner F, O'Sullivan N, Zink A, Chamel B, Khawam R, Coqueugniot E, Helmer D, Le Mort F, Perrin P, Gourichon L, Dutailly B,

Pálfi G, Coqueugniot H, Dutour O. 2015. Human tuberculosis predates domestication in ancient Syria. *Tuberculosis (Edinb)* **95**(Suppl 1)**:**S4–S12 http://dx.doi.org/10.1016/j.tube.2015.02.001.

18. Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, Greenblatt C, Donoghue H, Spigelman M, Brittain D. 2001. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis* **33:**305–311 http://dx.doi.org/10.1086/321886.

19. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM,et al. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* **31:**406–409.

20. van Soolingen D, van der Zanden AG, de Haas PE, Noordhoek GT, Kiers A, Foudraine NA, Portaels F, Kolk AH, Kremer K, van Embden JD. 1998. Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. *J Clin Microbiol* **36:**1840–1845.

21. Collins DM, Stephens DM. 1991. Identification of an insertion sequence, IS*1081*, in *Mycobacterium bovis*. *FEMS Microbiol Lett* **67:**11–15 http://dx.doi.org/10.1111/j.1574-6968.1991.tb04380.x.

22. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. 1991. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol* **29:**2578–2586.

23. Yeh RW, Ponce de Leon A, Agasino CB, Hahn JA, Daley CL, Hopewell PC, Small PM. 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. *J Infect Dis* **177:**1107–1111 http://dx.doi.org/10.1086/517406.

24. Fang Z, Morrison N, Watt B, Doig C, Forbes KJ. 1998. IS*6100* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol* **180:**2102–2109.

25. Achtman M. 1996. A surfeit of YATMs? *J Clin Microbiol* **34:**1870.

26. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* **35:**907–914.

27. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol* **10:**1057–1065 http://dx.doi.org/10.1111/j.1365-2958.1993.tb00976.x.

28. Gori A, Bandera A, Marchetti G, Degli Esposti A, Catozzi L, Nardi GP, Gazzola L, Ferrario G, van Embden JD, van Soolingen D, Moroni M, Franzetti F. 2005. Spoligotyping and *Mycobacterium tuberculosis*. *Emerg Infect Dis* **11:**1242–1248 http://dx.doi.org/10.3201/eid1108.040982.

29. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315:**1709–1712 http://dx.doi.org/10.1126/science.1138140.

30. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent V, Gicquel B, Tibayrenc M, Locht C, Supply P. 2001. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci USA* **98:**1901–1906 http://dx.doi.org/10.1073/pnas.98.4.1901.

31. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* **94:**9869–9874 http://dx.doi.org/10.1073/pnas.94.18.9869.

32. Brosch R, Pym AS, Gordon SV, Cole ST. 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9:**452–458 http://dx.doi.org/10.1016/S0966-842X(01)02131-X.

33. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA* **99:**3684–3689 http://dx.doi.org/10.1073/pnas.052548299.

34. Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, Laval F, Daffé M, Martin C, Brosch R, Guilhot C. 2014. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci USA* **111:**11491–11496 http://dx.doi.org/10.1073/pnas.1406693111.

35. Filliol I, Ferdinand S, Negroni L, Sola C, Rastogi N. 2000. Molecular typing of *Mycobacterium tuberculosis* based on variable number of tandem DNA repeats used alone and in association with spoligotyping. *J Clin Microbiol* **38:**2520–2524.

36. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS*1081*, IS*6100*, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol* **53:**680–689 http://dx.doi.org/10.1007/s002390010255.

37. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N. 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol* **3:**125–133 http://dx.doi.org/10.1016/S1567-1348(03)00011-X.

38. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S. 2008. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog* **4:**e1000160 http://dx.doi.org/10.1371/journal.ppat.1000160.

39. Baker L, Brown T, Maiden MC, Drobniewski F. 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* **10:**1568–1577 http://dx.doi.org/10.3201/eid1009.040046.

40. Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM. 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162:**1533–1543.

41. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* **188:**759–772 http://dx.doi.org/10.1128/JB.188.2.759-772.2006.

42. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN, Huynh LY, Fraser CM, Keim P. 2004. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA* **101:**13536–13541 http://dx.doi.org/10.1073/pnas.0403844101.

43. Galagan JE. 2014. Genomic insights into tuberculosis. *Nat Rev Genet* **15:**307–320 http://dx.doi.org/10.1038/nrg3664.

44. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:**537–544 http://dx.doi.org/10.1038/31159.

45. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, Fiette L, Orgeur M, Fabre M, Parmentier C, Frigui W, Simeone R, Boritsch EC, Debrie AS, Willery E, Walker D, Quail MA, Ma L, Bouchier C, Salvignol G, Sayes F, Cascioferro A, Seemann T, Barbe V, Locht C, Gutierrez MC, Leclerc C, Bentley SD, Stinear TP, Brisse S, Médigue C, Parkhill J, Cruveiller S, Brosch R. 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet* **45:**172–179 http://dx.doi.org/10.1038/ng.2517.

46. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, Supply P, Vincent V. 2005. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog* **1:**e5 http://dx.doi.org/10.1371/journal.ppat.0010005.

47. van Soolingen D, Hoogenboezem T, de Haas PE, Hermans PW, Koedam MA, Teppema KS, Brennan PJ, Besra GS, Portaels F, Top J, Schouls LM, van Embden JD. 1997. A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa. *Int J Syst Bacteriol* **47:**1236–1245 http://dx.doi.org/10.1099/00207713-47-4-1236.

48. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* **42:**498–503 http://dx.doi.org/10.1038/ng.590.

49. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet* **45:**1176–1182 http://dx.doi.org/10.1038/ng.2744.

50. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, Gadisa E, Kiros T, Habtamu M, Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg Infect Dis* **19:**460–463 http://dx.doi.org/10.3201/eid1903.120256.

51. Ho SY, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour* **11:**423–434 http://dx.doi.org/10.1111/j.1755-0998.2011.02988.x.

52. Drummond AJ, Rambaut A. 2007. BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7:**214 http://dx.doi.org/10.1186/1471-2148-7-214.

53. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22:**1185–1192 http://dx.doi.org/10.1093/molbev/msi103.

54. Bocquet-Appel JP. 2011. When the world's population took off: the springboard of the Neolithic Demographic Transition. *Science* **333:**560–561 http://dx.doi.org/10.1126/science.1208880.

55. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ, Campbell TJ, Majander K, Wilbur AK, Guichon RA, Wolfe Steadman DL, Cook DC, Niemann S, Behr MA, Zumarraga M, Bastida R, Huson D, Nieselt K, Young D, Parkhill J, Buikstra JE, Gagneux S, Stone AC, Krause J. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514:**494–497 http://dx.doi.org/10.1038/nature13591.

56. Bouwman AS, Kennedy SL, Müller R, Stephens RH, Holst M, Caffell AC, Roberts CA, Brown TA. 2012. Genotype of a historic strain of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* **109:**18511–18516 http://dx.doi.org/10.1073/pnas.1209444109.

57. Nicklisch N, Maixner F, Ganslmeier R, Friederich S, Dresely V, Meller H, Zink A, Alt KW. 2012. Rib lesions in skeletons from early Neolithic sites in Central Germany: on the trail of tuberculosis at the onset of agriculture. *Am J Phys Anthropol* **149:**391–404 http://dx.doi.org/10.1002/ajpa.22137.

58. Coros A, DeConno E, Derbyshire KM. 2008. IS*6100*, a *Mycobacterium tuberculosis* complex-specific insertion sequence, is also present in the genome of *Mycobacterium smegmatis*, suggestive of lateral gene transfer among mycobacterial species. *J Bacteriol* **190:**3408–3410 http://dx.doi.org/10.1128/JB.00009-08.

59. Müller R, Roberts CA, Brown TA. 2015. Complications in the study of ancient tuberculosis: non-specificity of IS*6100* PCRs. *Sci Technol Archeol Res* **1:**1–8 http://dx.doi.org/10.1179/2054892314Y.0000000002.

60. Minnikin DE, Minnikin SM, Parlett JH, Goodfellow M, Magnusson M. 1984. Mycolic acid patterns of some species of *Mycobacterium*. *Arch Microbiol* **139:**225–231 http://dx.doi.org/10.1007/BF00402005.

61. Minnikin DE, Parlett JH, Magnusson M, Ridell M, Lind A. 1984. Mycolic acid patterns of representatives of *Mycobacterium bovis* BCG. *J Gen Microbiol* **130:**2733–2736.

62. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog* **9:**e1003543 (Erratum: **9**[8]) http://dx.doi.org/10.1371/journal.ppat.1003543.

63. Kay GL, Sergeant MJ, Zhou Z, Chan JZ, Millard A, Quick J, Szikossy I, Pap I, Spigelman M, Loman NJ, Achtman M, Donoghue HD, Pallen MJ. 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* **6:**6717 http://dx.doi.org/10.1038/ncomms7717.

64. Moser I, Prodinger WM, Hotzel H, Greenwald R, Lyashchenko KP, Bakker D, Gomis D, Seidler T, Ellenberger C, Hetzel U, Wuennemann K, Moisson P. 2008. *Mycobacterium pinnipedii*: transmission from South American sea lion (*Otaria byronia*) to Bactrian camel (*Camelus bactrianus bactrianus*) and Malayan tapirs (*Tapirus indicus*). *Vet Microbiol* **127:**399–406 http://dx.doi.org/10.1016/j.vetmic.2007.08.028.

65. Kiers A, Klarenbeek A, Mendelts B, Van Soolingen D, Koëter G. 2008. Transmission of *Mycobacterium pinnipedii* to humans in a zoo with marine mammals. *Int J Tuberc Lung Dis* **12:**1469–1473.

66. Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, Simon S, Harris B, Atkin R, Doggett J, Mayes R, Keating L, Wheeler PR, Parkhill J, Barrell BG, Cole ST, Gordon SV, Hewinson RG. 2003. The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci USA* **100:**7877–7882 http://dx.doi.org/10.1073/pnas.1130426100.

67. Pym AS, Brodin P, Brosch R, Huerre M, Cole ST. 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and Mycobacterium microti. *Mol Microbiol* **46:**709–717 http://dx.doi.org/10.1046/j.1365-2958.2002.03237.x.

68. Mostowy S, Cousins D, Behr MA. 2004. Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. *J Bacteriol* **186:**104–109 http://dx.doi.org/10.1128/JB.186.1.104-109.2003.

69. Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, Gey van Pittius NC. 2010. Novel *Mycobacterium tuberculosis* complex pathogen, *M. mungi*. *Emerg Infect Dis* **16:**1296–1299 http://dx.doi.org/10.3201/eid1608.100314.

70. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S. 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* **10:**e1001387 http://dx.doi.org/10.1371/journal.pmed.1001387.

71. Mokrousov I. 2013. Insights into the origin, emergence, and current spread of a successful Russian clone of *Mycobacterium tuberculosis*. *Clin Microbiol Rev* **26:**342–360 http://dx.doi.org/10.1128/CMR.00087-12.

72. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rüsch-Gerdes S, Mokrousov I, Aleksic E, Allix-Béquec C, Antierens A, Augustynowicz-Kopeć E, Ballif M, Barletta F, et al. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis*

Beijing lineage. *Nat Genet* **47:**242–249 http://dx.doi.org/10.1038/ng.3195.

73. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M. 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445:**915–918 http://dx.doi.org/10.1038/nature05562.

74. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102:**15942–15947 http://dx.doi.org/10.1073/pnas.0507611102.

75. Laruelle M, Peyrouse S. 2009. Cross-border minorities as cultural and economic mediators between China and central Asia. *China Eurasia Forum Quarterly* **7:**93–119.

76. Wirth T. 2015. Massive lineage replacements and cryptic outbreaks of *Salmonella Typhi* in eastern and southern Africa. *Nat Genet* **47:**565–567 http://dx.doi.org/10.1038/ng.3318.

77. Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G, Shen X, Liu F, Gagneux S, Mei J, Lan R, Wan K, Gao Q. 2015. Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proc Natl Acad Sci USA* **112:**8136–8141 http://dx.doi.org/10.1073/pnas.1424063112.

78. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, Guthrie JL, Jamieson FB, Langlois-Klassen D, Long R, Nguyen D, Wobeser W, Feldman MW. 2011. Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade. *Proc Natl Acad Sci USA* **108:**6526–6531 http://dx.doi.org/10.1073/pnas.1016708108.

79. Lee RS, Radomski N, Proulx JF, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. 2015. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* **112:**13609–13614 http://dx.doi.org/10.1073/pnas.1507071112.

80. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* **13:**137–146 http://dx.doi.org/10.1016/S1473-3099(12)70277-3.

81. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. 2015. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun* **6:**7119 http://dx.doi.org/10.1038/ncomms8119.

82. Rado TA, Bates JH, Engel HW, Mankiewicz E, Murohashi T, Mizuguchi Y, Sula L. 1975. World Health Organization studies on bacteriophage typing of mycobacteria. Subdivision of the species *Mycobacterium tuberculosis*. *Am Rev Respir Dis* **111:**459–468.

83. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* **6:**e311 http://dx.doi.org/10.1371/journal.pbio.0060311.

84. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S. 2012. High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms. *PLoS One* **7:**e39855 http://dx.doi.org/10.1371/journal.pone.0039855.

85. Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* **4:**e7815 http://dx.doi.org/10.1371/journal.pone.0007815.

86. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* **5:**4812 http://dx.doi.org/10.1038/ncomms5812.

87. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SA, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van Soolingen D. 2013. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* **13:**110 http://dx.doi.org/10.1186/1471-2334-13-110.

88. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, Pym A, Mahayiddin AA, Chuchottaworn C, Sanne IM, Louw C, Boeree MJ, Hoelscher M, McHugh TD, Bateson AL, Hunt RD, Mwaigwisya S, Wright L, Gillespie SH, Bentley SD. 2013. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med* **1:**786–792 http://dx.doi.org/10.1016/S2213-2600(13)70231-5.

89. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, Flynn JL, Fortune SM. 2011. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* **43:**482–486 http://dx.doi.org/10.1038/ng.811.

90. Ewing G, Nicholls G, Rodrigo A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics* **168:**2407–2420 http://dx.doi.org/10.1534/genetics.104.030411.

91. Gray RR, Pybus OG, Salemi M. 2011. Measuring the temporal structure in serially-sampled phylogenies. *Methods Ecol Evol* **2:**437–445 http://dx.doi.org/10.1111/j.2041-210X.2011.00102.x.

92. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends Ecol Evol* **18:**481–488 http://dx.doi.org/10.1016/S0169-5347(03)00216-7.

93. Nübel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Zemlicková H, Leblois R, Wirth T, Jombart T, Balloux F, Witte W. 2010. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog* **6:**e1000855 http://dx.doi.org/10.1371/journal.ppat.1000855.

94. Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simões PM, Tristan A, Petersen A, Aziz M, Kiil K, Cirković I, Udo EE, del Campo R, Vuopio-Varkila J, Ahmad N, Tokajian S, Peters G, Schaumburg F, Olsson-Liljequist B, Givskov M, Driebe EE, Vigh HE, Shittu A, Ramdani-Bougessa N, Rasigade JP, Price LB, Vandenesch F, Larsen AR, Laurent F. 2014. Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. *MBio* **5:**e01044-14 http://dx.doi.org/10.1128/mBio.01044-14.

95. Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D, Andrews BJ. 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* **1:**183–186 http://dx.doi.org/10.1534/g3.111.000406.

96. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* **109:**E2774–E2783 http://dx.doi.org/10.1073/pnas.1210309109.

97. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S. 2011. *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci USA* **108:**5033–5038 http://dx.doi.org/10.1073/pnas.1018444108.

98. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. 2013. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc Natl Acad Sci USA* **110:**13880–13885 http://dx.doi.org/10.1073/pnas.1304681110.

99. Linz B, Windsor HM, McGraw JJ, Hansen LM, Gajewski JP, Tomsho LP, Hake CM, Solnick JV, Schuster SC, Marshall BJ. 2014. A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques. *Nat Commun* **5:**4165 http://dx.doi.org/10.1038/ncomms5165.

100. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. *Trends Ecol Evol* **30:**306–313 http://dx.doi.org/10.1016/j.tree.2015.03.009.

101. Ho SY, Larson G. 2006. Molecular clocks: when times are a-changin'. *Trends Genet* **22:**79–83 http://dx.doi.org/10.1016/j.tig.2005.11.006.

102. Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Mol Ecol* **20:**3087–3101 http://dx.doi.org/10.1111/j.1365-294X.2011.05178.x.

103. Bandelt HJ. 2008. Clock debate: when times are a-changin': time dependency of molecular rate estimates: tempest in a teacup. *Heredity (Edinb)* **100:**1–2 http://dx.doi.org/10.1038/sj.hdy.6801054.

104. Burridge CP, Craw D, Fletcher D, Waters JM. 2008. Geological dates and molecular rates: fish DNA sheds light on time dependency. *Mol Biol Evol* **25:**624–633 http://dx.doi.org/10.1093/molbev/msm271.

105. Ho SY, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol* **22:**1561–1568 http://dx.doi.org/10.1093/molbev/msi145.

106. Penny D. 2005. Evolutionary biology: relativity for molecular clocks. *Nature* **436:**183–184 http://dx.doi.org/10.1038/436183a.

107. Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, Cheng J, Wang W, Wang J, Qian W, Li D, Wang L. 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS One* **3:**e4053 http://dx.doi.org/10.1371/journal.pone.0004053.

108. Jamrozy D, Kallonen T. 2015. Looking at Beijing's skyline. *Nat Rev Microbiol* **13:**528 http://dx.doi.org/10.1038/nrmicro3536.

109. Rodrigo A, Bertels F, Heled J, Noder R, Shearman H, Tsai P. 2008. The perils of plenty: what are we going to do with all these genes? *Philos Trans R Soc Lond B Biol Sci* **363:**3893–3902 http://dx.doi.org/10.1098/rstb.2008.0173.

110. Martincorena I, Seshasayee AS, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485:**95–98 http://dx.doi.org/10.1038/nature10995.

111. Dos Vultos T, Mestre O, Tonjum T, Gicquel B. 2009. DNA repair in *Mycobacterium tuberculosis* revisited. *FEMS Microbiol Rev* **33:**471–487 http://dx.doi.org/10.1111/j.1574-6976.2009.00170.x.

112. Boshoff HI, Reed MB, Barry CE III, Mizrahi V. 2003. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Cell* **113:**183–193 http://dx.doi.org/10.1016/S0092-8674(03)00270-8.

113. Ebrahimi-Rad M, Bifani P, Martin C, Kremer K, Samper S, Rauzier J, Kreiswirth B, Blazquez J, Jouan M, van Soolingen D, Gicquel B. 2003. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* **9:**838–845 http://dx.doi.org/10.3201/eid0907.020803.

114. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* **45:**784–790 http://dx.doi.org/10.1038/ng.2656.

115. Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, Taddei F. 2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* **291:**2606–2608 http://dx.doi.org/10.1126/science.1056421.

116. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. 2000. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288:**1251–1254 http://dx.doi.org/10.1126/science.288.5469.1251.

117. Mizrahi V, Andersen SJ. 1998. DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? *Mol Microbiol* **29:**1331–1339 http://dx.doi.org/10.1046/j.1365-2958.1998.01038.x.

118. Ng VH, Cox JS, Sousa AO, MacMicking JD, McKinney JD. 2004. Role of KatG catalase-peroxidase in mycobacterial pathogenesis: countering the phagocyte oxidative burst. *Mol Microbiol* **52:**1291–1302 http://dx.doi.org/10.1111/j.1365-2958.2004.04078.x.

119. Sassetti CM, Rubin EJ. 2003. Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci USA* **100:**12989–12994 http://dx.doi.org/10.1073/pnas.2134250100.

120. Orlando L, Cooper A. 2014. Using ancient DNA to understand evolutionary and ecological processes. *Annu Rev Ecol Evol Syst* **45:**573–598 http://dx.doi.org/10.1146/annurev-ecolsys-120213-091712.

121. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444:**330–336 http://dx.doi.org/10.1038/nature05336.

122. Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA damage. *Cold Spring Harb Perspect Biol* **5:**a012567. http://dx.doi.org/10.1101/cshperspect.a012567.

123. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Hensen JD, Marques-Bonet T, et al. 2010. A draft sequence of the Neanderthal genome. *Science* **328:**710–722 http://dx.doi.org/10.1126/science.1188021.

124. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TF, Slatkin M, Hublin JJ, Reich D, Kelso J, Viola TB, Pääbo S. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514:**445–449 http://dx.doi.org/10.1038/nature13810.

125. Pääbo S. 2015. The diverse origins of the human gene pool. *Nat Rev Genet* **16:**313–314 http://dx.doi.org/10.1038/nrg3954.

126. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505:**43–49 http://dx.doi.org/10.1038/nature12886.

127. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338:**222–226 http://dx.doi.org/10.1126/science.1224344.

128. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJ, Herring DA, Bauer P, Poinar HN, Krause J. 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478:**506–510 http://dx.doi.org/10.1038/nature10549.

129. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren KG, Pedersen AG, Schubert M, Van Dam A, Kapel CM, Nielsen HB, Brunak S, Avetisyan P, Epimakhov A, Khalyapin MV, Gnuni A, Kriiska A, Lasak I, Metspalu M, Moiseyev V, Gromov A, Pokutta D, Saag L, Varul L, Yepiskoposyan L, Sicheritz-Pontén T, Foley RA, Lahr MM, Nielsen R, Kristiansen K, Willerslev E. 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163:**571–582 http://dx.doi.org/10.1016/j.cell.2015.10.009.

130. Pálfi G, Dutour O, Perrin P, Sola C, Zink A. 2015. Tuberculosis in evolution. *Tuberculosis (Edinb)* **95**(Suppl 1)**:**S1–S3 http://dx.doi.org/10.1016/j.tube.2015.04.007.

131. Pálfi G, Maixner F, Maczel M, Molnár E, Pósa A, Kristóf LA, Marcsik A, Balázs J, Masson M, Paja L, Palkó A, Szentgyörgyi R, Nerlich A, Zink A, Dutour O. 2015. Unusual spinal tuberculosis in an Avar Age skeleton (Csongrád-Felgyő, Ürmös-tanya, Hungary): A morphological and biomolecular study. *Tuberculosis (Edinb)* **95**(Suppl 1)**:**S29–S34 http://dx.doi.org/10.1016/j.tube.2015.02.033.

132. Pósa A, Maixner F, Mende BG, Köhler K, Osztás A, Sola C, Dutour O, Masson M, Molnár E, Pálfi G, Zink A. 2015. Tuberculosis in Late Neolithic-Early Copper Age human skeletal remains from Hungary. *Tuberculosis(Edinb)* **95**(Suppl 1)**:**S18–S22 http://dx.doi.org/10.1016/j.tube.2015.02.011.

133. Lee OY, Wu HH, Besra GS, Rothschild BM, Spigelman M, Hershkovitz I, Bar-Gal GK, Donoghue HD, Minnikin DE. 2015. Lipid biomarkers provide evolutionary signposts for the oldest known cases of tuberculosis. *Tuberculosis (Edinb)* **95**(Suppl 1)**:**S127–S132 http://dx.doi.org/10.1016/j.tube.2015.02.013.

Chapitre 2.

Les mutations compensatoires pilotent l'épidémie de souches MDR en Asie Centrale

Classification: Biological Sciences, Genetics

Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia

Authors: Matthias Merker[1,2*], Maxime Barbier[3,4*], Helen Cox[5], Jean-Philippe Rasigade[3,4,6], Silke Feuerriegel[1,2], Thomas A. Kohl[1,2], Roland Diel[7], Sonia Borrell[8,9], Sebastien Gagneux[8,9], Vladyslav Nikolayevskyy[10,11], Sönke Andres[12], Ulrich Nübel[13,14], Philip Supply[15,16,17,18], Thierry Wirth[3,4], Stefan Niemann[1,2#]

Affiliations:

[1]Molecular and Experimental Mycobacteriology, Research Center Borstel, Germany.

[2]German Center for Infection Research, Borstel site, Germany.

[3]Laboratoire Biologie Intégrative des Populations, Evolution Moléculaire, Ecole Pratique des Hautes Etudes, PSL University, Paris, France.

[4]Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Université Pierre et Marie Curie, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France.

[5]Division of Medical Microbiology and Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa.

[6]CIRI INSERM U1111, University of Lyon, Lyon, France.

[7] Institute for Epidemiology, Schleswig-Holstein University Hospital, Kiel, Germany

[8] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland.

[9] University of Basel, Basel, Switzerland.

[10] Imperial College London, United Kingdom.

[11] Public Health England, London, United Kingdom

[12] Division of Mycobacteriology (National Tuberculosis Reference Laboratory), Research Center Borstel, Borstel. Germany

[13]Microbial Genome Research, Leibniz-Institut DSMZ- Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig, Germany.

[14] German Center for Infection Research, Braunschweig site, Germany.

[15] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 8204 - CIIL - Centre d'Infection et d'Immunité de Lille, F-59000 Lille, France.

[16] Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 8204, Center for Infection and Immunity of Lille, Lille, France.

[17] Université Lille Nord de France, Center for Infection and Immunity of Lille, Lille, France.

[18] Institut Pasteur de Lille, Center for Infection and Immunity of Lille, Lille, France.

* contributed equally

Correspondence to: Prof. Dr. Stefan Niemann, Head Molecular Mycobacteriology Group, National Reference Center for Mycobacteria, Deputy Head Priority Area Infections

Forschungszentrum Borstel

Parkallee 1, 23845 Borstel, Germany

E-Mail: sniemann@fz-borstel.de

Phone: 0049-4537-1887620

Fax: 0049-4537-1882091

Keywords: Mycobacterium tuberculosis, multidrug resistance, compensatory evolution

Abstract

Multidrug resistant tuberculosis (MDR-TB, caused by isoniazid and rifampicin resistant strains) is a globally spreading, life-threatening infectious disease, whose incidence culminates in Central Asia and the Russian Federation. The contributions of bacterial genetic and/or programmatic TB management-related factors to this MDR-TB epidemic remain largely unclear. Here, we utilized whole genome sequencing (WGS) and Bayesian statistics to analyze the evolutionary history, emergence of resistance and transmission networks of 277 MDR *Mycobacterium tuberculosis* complex (MTBC) strains from Karakalpakstan, Uzbekistan, collected between 2001 and 2006. Dating analyses revealed that one particular MTBC clone (termed Central Asian outbreak, CAO) with resistance mediating mutations to eight different anti-TB drugs existed prior the worldwide introduction of standardized WHO-endorsed directly observed treatment, short-course (DOTS). DOTS implementation in Karakalpakstan in 1998 likely selected for these CAO-strains, comprising 75% of sampled MDR-TB strains in 2005/2006. Transmission success and resistance development was linked to mutations in genes coding for the RNA-polymerase subunits *rpoA*, *rpoB*, and *rpoC* postulated to compensate fitness deficits associated with acquisition of rifampicin resistance. Combined analysis with a previous dataset of 428 genomes from MDR-MTBC strains revealed the presence of CAO-strains also in Samara, Russia (2008-2010), and suggests that MDR-TB in Russia and Central Asia is driven by three clonal outbreaks causing 70% of all MDR-TB cases in both settings. The genetic make-up of these outbreak clades threatens the success of both empirical and standardized guideline driven MDR-TB therapies, including the newly WHO-endorsed short MDR-TB regimen in Uzbekistan.

Significance statement

Multidrug-resistant tuberculosis (MDR-TB) is considered as public health crisis. However, little is known about bacterial evolutionary trajectories in high MDR-TB incidence settings, the temporal acquisition of drug-resistances and how programmatic factors influence bacterial population dynamics. We focus on a Central Asian setting by integrating whole genome-based transmission analysis and Bayesian statistics to recap the evolutionary history of MDR-TB strains in Central Asia. The most transmissible MDR-clade is likely influenced by programmatic and social economic changes. Evolutionary success is linked with the fixation of multiple drug-resistance related and fitness-related mutations already in the 1980s, prior to the dissolution of the Soviet Union. Importantly, this initial phase was followed by selective expansion upon implementation of DOTS therapy without appropriate drug resistance diagnostics.

Introduction

Multidrug-resistant tuberculosis (MDR-TB), caused by *Mycobacterium tuberculosis* complex (MTBC) strains that are resistant to the first-line drugs isoniazid and rifampicin, represent a threat to global TB control. Barely 20% of the estimated annual 480,000 new MDR-TB patients have access to adequate second-line treatment regimens. The majority of undiagnosed or ineffectively treated MDR-TB patients continue to transmit their infection and suffer high mortality (1).

Based on early observations that the acquisition of drug resistance could lead to reduced bacterial fitness (2) it was hypothesized that drug-resistant MTBC-strains had a reduced capacity to transmit, and would not widely disseminate in the general population (3–7). This optimistic scenario has been invalidated by the now abundant evidence for transmission of MDR and extensively drug-resistant MTBC-strains (XDR-TB; MDR-TB additionally resistant to at least one fluoroquinolone and one aminoglycoside) in healthcare and community settings (3, 8–11). In former Soviet Union countries, which experience the highest MDR-TB rates worldwide, the expansion of drug-resistant MTBC-clones is thought be promoted by interrupted drug supplies, inadequate implementation of regimens, lack of infection control and erratic treatment in prison settings (12, 13). Continued transmission is thought to be aided by the co-selection of mutations in the bacterial population that compensate for a fitness cost (e.g. growth deficit) associated particularly with the acquisition of rifampicin resistance mediating mutations (3, 7–11). The compensatory mechanism for rifampicin resistant MTBC strains is proposed to be associated with structural changes in the RNA-polymerase subunits *RpoA*, *RpoB*, and *RpoC* that increase transcriptional activity and as a consequence enhance the growth rate (11). However, the impact of these bacterial genetic factors on the epidemiological success of MDR-MTBC strains and implications for current and upcoming MDR-TB treatment strategies remain unexplored.

We utilized whole genome sequencing (WGS) to retrace the longitudinal transmission and evolution of MTBC-strains towards MDR/pre-XDR/XDR geno- and phenotypes in Karakalpakstan, Uzbekistan. In this high MDR-TB incidence setting, the proportion of MDR-TB among new TB-patients increased from 13% in 2001 to 23% in 2014 despite the local introduction of the World Health Organization recommended

DOTS strategy in 1998 and an initially limited MDR-TB treatment program in 2003 (14, 15). We expanded our analyses by including a WGS dataset of MDR-MTBC strains from Samara, Russia (2008-2010) (13) to investigate clonal relatedness, resistance and compensatory evolution in both settings.

Methods

Study populations

Karakalpakstan, Uzbekistan

A total of 277 MDR-MTBC strains derived from two separate cohorts were sequenced. The first cohort comprised 86% (49/57) of MDR-MTBC strains from a cross-sectional drug resistance survey conducted in four districts in Karakalpakstan, Uzbekistan between 2001-2002 (16). An additional 228 strains were obtained from TB-patients enrolled for second-line treatment in the MDR-TB treatment program from 2003 to 2006. These strains represented 76% (228/300) of all MDR-TB cases diagnosed over the period. While the MDR-TB treatment program covered two of the four districts included in the initial drug resistance survey, the majority of strains from both cohorts, 69% and 64% respectively, were obtained from patients residing in the same main city of Nukus (Table S1).

Samara, Russia

To set the MDR-MTBC strains from Karakalpakstan into a broader geographical perspective, raw sequencing data of 428 MDR-MTBC strains from a published cross-sectional prospective study in Samara, Russia from 2008-2010 (13) were processed as described below and included into a composite MDR-MTBC dataset.

Drug susceptibility testing

Drug susceptibility testing (DST) was performed for five first-line drugs (isoniazid, rifampicin, ethambutol, streptomycin, pyrazinamide), and three second-line drugs (ofloxacin, capreomycin and prothionamide) for cohort 1, and six second-line drugs for cohort 2 (capreomycin, amikacin, ofloxacin, ethionamide, para-aminosalicylic acid and cycloserine) by the reference laboratory in Borstel, Germany as described previously (17).

Whole genome sequencing

Whole genome sequencing (WGS) was performed with Illumina Technology (MiSeq and HiSeq 2500) using Nextera XT library preparation kits as instructed by the manufacturer (Illumina, San Diego, CA, USA). Fastq files (raw sequencing data) were submitted to the European nucleotide archive (see additional data for accession numbers). Obtained reads were mapped to the *M. tuberculosis* H37Rv reference genome (GenBank ID: NC_000962.3) with BWA (18). Alignments were refined with GATK (19) and Samtools (20) toolkits with regard to base quality re-calibration and alignment corrections for possible PCR artefact. For variant detection in mapped reads, we employed Samtools and custom perl scripts to filter for thresholds of a minimum coverage of a total of 4 reads in both forward and reverse orientation, 4 reads calling the allele with at least a phred score of 20, and 75% allele frequency. In the combined datasets, we allowed a maximum of 5% of all samples to fail the above mentioned threshold criteria in individual genome positions to compensate for coverage fluctuations in certain genome regions; in these cases, the majority allele was considered. Regions annotated as 'repetitive' elements (e.g. PPE and PE-PGRS gene families), insertions and deletions (InDels), and consecutive variants in a 12 bp window (putative artefacts flanking InDels) were excluded. Additionally, 32 resistance associated target regions (see additional data) were excluded for phylogenetic reconstructions. The remaining single nucleotide polymorphisms (SNPs) were considered as valid and used for concatenated sequence alignments.

Transmission index

Based on the distance matrix (SNP distances), we further determined for every isolate the number of isolates that were in a range of 10 SNPs or less (in the following referred to as "transmission index"). This 10 SNP-threshold was used to infer the number of recently linked cases, as considered within a 10-year time period, based on previous convergent estimates of MTBC genome evolution rate of ≈ 0.5 SNPs/genome/year in inter-human transmission chains and in macaque infection models (21–24),

Genotypic drug resistance prediction

Mutations (small deletions and SNPs) in 32 resistance associated target regions were considered for a molecular resistance prediction to 13 first- and second-line drugs (additional data). Mutations in genes coding for the RNA-Polymerase subunits *rpoA*, *rpoB* (excluding resistance mediating mutations), and *rpoC* were reported as putative fitness compensating (e.g. in vitro growth enhancing) variants for rifampicin resistant strains. A detailed overview of all mutations considered as genotypic resistance marker is given as additional data. Mutations that were not clearly linked to phenotypic drug resistance were reported as genotypic non wild type and were not considered as genotypic resistance markers. When no mutation (or synonymous, silent mutations) was detected in any of the defined drug relevant target regions the isolate was considered to be phenotypically susceptible.

Phylogenetic inference (maximum likelihood)

We used jModelTest v2.1 and Akaike and Bayesian Information Criterion (AIC and BIC) to find an appropriate substitution model for phylogenetic reconstructions based on the concatenated sequence alignments (Table S2). Maximum likelihood trees were calculated with FastTree 2.1.9 (double precision for short branch lengths) (25) using a general time reversible (GTR) nucleotide substitution model (best model according to AIC and second best model according to BIC), 1,000 resamplings and Gamma20 likelihood optimization to account for evolutionary rate heterogeneity among sites. The consensus tree was rooted with the "midpoint root" option in FigTree and nodes were arranged in increasing order. Polymorphisms considered as drug resistance marker (see above) and putative compensatory variants were analyzed individually and mapped on the phylogenetic tree to define resistance patterns of identified phylogenetic subgroups.

Molecular clock model

In order to compute a time scaled phylogeny and employ the Bayesian skyline model (see below) for the identified Central Asian outbreak (CAO) clade we sought to define an appropriate molecular clock model (strict versus relaxed clock) and a mutation rate estimate. Due to the restricted sampling timeframe of the Karakalpakstan dataset (2001-2006) we extended the dataset for the model selection process with CAO strains from Samara (2008-2010) and 'historical' CAO strains isolated from MDR-TB patients in Germany

(1995-2000) thus allowing for a more confident mutation rate estimate. The strength of the temporal signal in the combined dataset, assessed by the correlation of sampling year and root-to-tip distance, was investigated with TempEst v1.5 (26). Regression analysis was based on residual mean squares, using a rooted ML tree (PhyML, GTR substitution model, 100 bootstraps), R-square and adjusted *P*-value are reported. For the comparison of different Bayesian phylogenetic models we used path sampling with an alpha of 0.3, 50% burn-in and 15 million iterations (resulting in mean ESS values >100), marginal likelihood estimates were calculated with BEAST v2.4.2 (27), and Δ marginal L estimates are reported relative to the best model.

First, we employed a strict molecular clock fixed to 1 x 10^{-7} substitutions per site per year as reported previously (21–23) without tip dating, a strict molecular clock with tip dating and a relaxed molecular clock with tip dating. BEAST templates were created with BEAUti v2 applying a coalescent constant size demographic model, a GTR nucleotide substitution model, a chain length of 300 million (10% burn-in) and sampling of 5,000 traces/trees.

Second, we ran different demographic models (i.e. coalescent constant size, exponential, and Bayesian skyline) under a relaxed molecular clock using tip dates and the same parameters for the site model and Markov-Chain-Monte-Carlo (MCMC) as described above. Inspection of BEAST log files with Tracer v1.6 showed an adequate mixing of the Markov chains and all parameters were observed with an effective sample size (ESS) in the hundreds, suggesting an adequate number of effectively independent draws from the posterior sample and thus sufficient statistical support.

Bayesian Skyline Plot

Changes of the effective population size of the CAO clade in Karakalpakstan over the last four decades were calculated with a Bayesian skyline plot using BEAST v2.4.2 (27) using a tip date approach with a strict molecular clock model of 0.94 x 10^{-7} substitutions per site per year (best model according to path sampling results, see above), and a GTR nucleotide substitution model. We further used a random starting tree, a chain length of 300 million (10% burn-in) and collected 5,000 traces/trees. Again adequate mixing

of the Markov chains and ESS values in the hundreds were observed. A maximum clade credibility genealogy was calculated with TreeAnnotator v2.

Impact of resistance-conferring and compensatory mutations on transmission success

We used multiple linear regression to examine the respective contributions of antimicrobial resistance and putative fitness cost-compensating mutations to the transmission success of tuberculosis. To take transmission duration into account, we computed, for each isolate and each period length *T* in years (from 1 to 40y before sampling), a transmission success score defined as the number of isolates distant of less than *T* SNPs, divided by *T*. This approach relied on the following rationale: based on MTBC evolution rate of 0.5 mutation per genome per year, the relation between evolution time and SNP divergence is such that a cluster with at most *N* SNPs of difference is expected to have evolved for approximately *N* years. Thus, transmission success score over *T* years could be interpreted as the size of the transmission network divided by its evolution time, hence as the average yearly increase of the network size. For each period *T*, the transmission success score was regressed on the number of resistance mutations and on the presence of putative compensatory mutations. The regression coefficients with 95% confidence intervals were computed and plotted against *T* to identify maxima, that is, time periods when the transmission success was maximally influenced by either resistance-conferring or –compensating mutations. These analyses were conducted independently on outbreak strains of the Beijing-CAO clade in the Karakalpakstan cohort and of the Beijing-A clade in the Samara cohort.

Statistical analyses.

Comparison of proportions between cohorts was performed using Chi-squared analysis (mid-P exact) or Fisher's exact test, while comparison of median age was performed using the Mann-Whitney test. *P*-values for pairwise comparisons of subgroups regarding pairwise genetic distances, number of resistant DST results and number of resistance related mutations were calculated with an unpaired t-test (Welch correction) or a t-test according to the result of the variances comparison using a F-test. Boxplot, bubble plots and density plots have been performed on R.

Results

Study population and MTBC phenotypic resistance (Karakalpakstan, Uzbekistan)

Despite differences in sampling for cohort 1 and cohort 2 (see methods), patients showed similar age, sex distributions, and proportion of residence in Nukus, the main city in Karakalpakstan (Uzbekistan) (Table S1). While the majority of strains from both cohorts were phenotypically resistant to additional first-line TB drugs (i.e. beyond rifampicin and isoniazid), combined resistance to all 5 first-line drugs was significantly greater in cohort 2 (47% in cohort 2 compared to 14% in cohort 1, $P < 0.0001$). The same was true for resistance to the second-line injectable drug capreomycin (23% in cohort 2 compared to 2% in cohort 1, $P = 0.0001$) (Table S1). This finding was surprising as the isolates from cohort 2 patients - who were treated with individualized second-line regimens predominately comprising ofloxacin as the fluoroquinolone and capreomycin as the second-line injectable - were all obtained before the initiation of their treatment. In addition, there was no formal MDR-TB treatment program in Karakalpakstan prior to 2003. These elements imply that the higher rate of resistance to capreomycin was attributable to infection by already resistant strains (i.e. to primary resistance).

MTBC population structure and transmission rates

Utilizing WGS, we determined 6,979 single nucleotide polymorphisms (SNPs) plus 537 variants located in 34 genes and upstream regions associated with drug resistance and bacterial fitness (additional data). The corresponding phylogeny revealed a dominant subgroup comprising 173/277 (62.5%) closely related strains within the Beijing-genotype (alias MTBC lineage 2) (Fig. 1). This group, termed Central Asian Outbreak (CAO), showed a highly restricted genetic diversity (median pairwise distance of 21 SNPs, IQR 13-25) and was differentiated from a set of more diverse strains by 38 specific SNPs (Fig. S1, additional data). The proportion of CAO-strains was similar between 2001-02 and 2003-04 (49% and 52% respectively), but increased to 76% in 2005-06 ($P < 0.01$). Over the same time periods, the proportions of other strain types remained stable or decreased (Fig. S2).

We then sized transmission networks (measured by transmission indexes, see methods) supposed to reflect human-to-human transmission over the last ~10 years based on a maximum of 10 differentiating SNPs

between two strains. Transmission rates varied, even among closely related outbreak strains (Fig. 1). Beijing-CAO-strains formed particularly large transmission networks (>50 strains/patients; Fig. 1); 96.0% (166/173) of all Beijing-CAO strains were associated with recent transmission (i.e. transmission index ≥1), versus 48.4% (31/64) of non-CAO Beijing strains (P<0.0001) and 57.5% (23/40) of non-Beijing strains (P<0.0001) (additional data). In addition the large CAO transmission network exhibited higher levels of drug resistance relative to non-Beijing strains, as reflected by the larger number of drugs for which phenotypic (P=0.0079) and genotypic drug resistance (P=0.0048) was detected Fig. S3).

Evolutionary history of CAO strains in Karakalpakstan

In order to gain more detailed insights into the emergence of resistance mutations in the evolutionary history of the CAO clade, we sought to employ a Bayesian phylogenetic analysis for a temporal calibration of the CAO phylogeny and an estimation of the mutation rate. Using an extended collection of more diverse CAO strains (n=220) from different settings (see methods) we initially compensated for the restricted sampling time frame of the Karakalpakstan dataset (2001-2006). A linear regression analysis showed correlation between sampling year and root-to-tip distance and a moderate temporal signal (P=0.00039, R^2= 5.2%, Fig. S4), allowing for further estimation of CAO mutation rates and evaluation of molecular clock models using Bayesian statistics. Based on the marginal L estimates collected by path sampling, we found a strict molecular clock with tip dates to be most appropriate (Table S3). Mutation rate estimates (under a relaxed clock model) ranged on average from 0.88 to 0.96 x 10^{-7} substitutions per site per year (s/s/y), depending on the demographic model, in favor for the Bayesian skyline model with mutation rate of 0.94 x 10^{-7} (s/s/y) (95% HPD 0.72-1.15 x 10^{-7} (s/s/y)) (Table S3).

We then employed the Bayesian skyline model with a strict molecular clock set to 0.94 x 10^{-7} (s/s/y) specifically for the CAO clade from Karakalpakstan (n=173). We determined that the most recent common ancestor (MRCA) of the CAO-clade emerged around 1976 (95% highest posterior density (HPD) 1969-1982). The MRCA already exhibited a streptomycin resistance mutation (*rpsL* K43R) (Fig. 2), and acquired isoniazid resistance (*katG* S315T) in 1977 (95% HPD 1973-1983). The CAO-population size then rose contemporaneously with multiple events of rifampicin, ethambutol, ethionamide, and para-aminosalicylic

acid resistance acquisition in different branches (Fig. 2). As an illustration, the most frequent CAO-clone (upper clade in Fig. 2) acquired ethambutol and ethionamide resistance mutations (*embB* M306V, *ethA* T314I) around 1984 (95% HPD 1982-1989), and an MDR-genotype (*rpoB* S450L) around 1986 (95% HPD 1985-1992). The effective population size reached a plateau before fixation of mutations in the *ribD* promoter region (leading to para-aminosalicylic acid resistance) and *rpoC* N698S, putatively enhancing its fitness around 1990 (95% HPD 1989-1994) (Fig. 2). Independent fixation of pyrazinamide (*pncA* Q10P and I133T) and kanamycin (*eis* -12 g/a) resistance-associated mutations was detected in 1992 and 1991 (both with 95% HPD rounded to 1991-1996) (Fig. 2).

Interestingly, the implementation of the systematic DOTS-program in Karakalpakstan in 1998 coincided with a second effective population size increase (Fig. 2). At that time, distinct CAO-subgroups already exhibited pre-XDR (in this context MDR plus kanamycin resistance) resistance profiles, mediating resistance to as many as eight different anti-TB drugs. Of note, only a single strain was identified as harboring a *gyrA* mutation (A90V), associated to fluoroquinolone resistance (additional data). At the end of the study period in 2006 we observed a pre-XDR rate among CAO strains of 52.0% (90/173), compared to 35.9% (23/64) among other Beijing strains (P=0.03) and compared to 42.5% (17/40) among non-Beijing strains (P=0.30) (additional data).

Impact of compensatory variants on transmission networks

Overall, 62.1% (172/277) of all MDR-MTBC strains carried putative compensatory mutations (11, 13) in *rpoA* (n=7), *rpoC* (n=126) and *rpoB* (n=43) (additional data). These mutations were almost completely mutually exclusive, as only 4/172 strains harbored variants in more than one RNA polymerase-encoding gene. While mutations in *rpoA* and *rpoB* were equally distributed between Beijing-CAO strains and other non-outbreak Beijing strains, CAO-strains had more *rpoC* variants (56% vs 28%, P=0.003) (Table S4). The mean number of resistance mutations was higher among strains carrying compensatory mutations (Fig. 3A), 4.77 vs 3.35 mutations (two-sample t-test $P=1.2\text{x}10^{-10}$). Notably, strains with compensatory mutations also showed larger transmission indexes than strains presenting no compensatory mutation, 37.16 vs 9.22 (Welch two-sample t-test $P<2.2\text{x}10^{-16}$) (Fig. 3B). CAO-strains with compensatory mutation also had more

resistance-conferring mutations than CAO-strains lacking such mutation (ANOVA, Tukey multiple comparisons of means *P* adj=0.0000012). There was no difference observed for the means of resistance-conferring mutations amon non-CAO strains; compensatory mutation present vs. absent (*P* adj= 0.1978623) (Fig. 3C).

Regression-based analyses of transmission success scores in the Beijing-CAO clade confirmed that the presence of compensatory mutations was strongly associated with cluster sizes independent of the accumulation of resistance mutations (Fig. 4). This pattern was mostly observed for clusters initiated in the late 1980s and the 1990s.

Combined analysis of MDR-TB cohorts from Karakalpakstan and Samara (2001-2010)

To place our analyses in a broader phylogenetic and geographic context, we combined our Karakalpakstan genome set with previously published genomes of 428 MDR-MTBC isolates from Samara (13), a Russian region located ~1,700 km from Nukus, Karakalpakstan. This analysis showed that Beijing-CAO strains accounted for the third largest strain clade in Samara (13). Conversely, the second largest clade in Samara, termed Beijing clade B according to Casali et al (13, 28), or European/Russian W148 (29), was represented in Karakalpakstan by a minor clade (Fig. 5). Considering a third Beijing clade (termed clade A) restricted to Samara (13), three major Beijing outbreak clades accounted for 69.6% (491/705) of the MDR-TB cases in both regions.

The three Beijing clades (A, B, and CAO) in Samara and Karakalpakstan had more drug resistance conferring mutations (in addition to isoniazid and rifampicin resistance) with means of 5.0 (SEM 0.07), 4.2 (SEM 0.18), and 4.7 (SEM 0.11), respectively (Fig. S5), than compared to only 3.6 (SEM 0.20) additional genotypic drug resistances ($P < 0.0001$, $P = 0.0143$, $P < 0.0001$) for other Beijing strains in both settings. Strains belonging to other MTBC genotypes (mainly lineage 4 subgroups) were found with a mean of 2.6 (SEM 0.20) additional drug resistance mediating mutations, lower than any Beijing-associated group ($P \leq 0.0009$) (Fig. S5).

Similar to Karakalpakstan, MDR-MTBC strains from Samara with compensatory mutations also accumulated more resistance-associated mutations (4.57 vs 2.30 mutations per genome; two-sample t-test

$P<2.2x10^{-16}$) and had higher transmission indexes (50.32 vs 0.46; Welch two-sample t-test $P<2.2x10^{-16}$) compared to strains lacking compensatory mutations (Fig. S6).

The impact of resistance conferring and compensatory mutations on the transmission success score in Beijing-A clade from Samara (Fig S7) was strikingly similar to the one observed in CAO strains from Karakalpakstan. The presence of compensatory mutations, but not the accumulation of resistance mutations, was significantly and independently associated with network size in clusters originating in the 1980s and 1990s, with a maximum influence found in clusters starting in the late 1990s.

Critically, the high proportions of strains detected in both settings with pre-XDR and XDR resistance profiles among the three major Beijing clades (clade A, 96%; clade B, 62%; clade CAO, 50%; Table S5, Figure 6) reveal the low proportion of patients that are or would be eligible to receive the newly WHO endorsed short MDR-TB regimen. As per definition of the WHO exclusion criteria, e.g. any confirmed or suspected resistance to one drug (except isoniazid) in the short regimen, only 0.5% (1/191 in Karakalpakstan) and 2.7% (8/300 in Samara) of the patients infected with either a Beijing clade A, B or CAO strain would benefit from a shortened MDR-TB therapy (additional data).

Discussion

Using WGS combined with Bayesian and phylogenetic analyses, we reveal the evolutionary history and recent clonal expansion of the dominatant MDR/pre-XDR MTBC clade in Karakalpakstan, Uzbekistan, termed the Central Asian outbreak (CAO). Strikingly, CAO-strains were also found also in Samara, Russia, and vice versa strains belonging to the second largest clade in Samara (Beijing clade B, i.e. European/Russian W148 (13, 29)) were identified in Karakalpakstan, suggesting that the MDR-TB epidemic in this world region is driven by few outbreak clades. During the three last decades, these strains gradually accumulated resistance to multiple anti-TB drugs that largely escaped phenotypic and molecular diagnostics, and reduce treatment options to a restricted set of drugs that often cause severe side effects. In addition, our results suggest that compensatory mutations (in RNA-polymerase subunit coding genes) that

are proposed to ameliorate growth deficits in rifampicin resistant strains in vitro are also crucial in a global epidemiological context allowing MDR and pre-XDR strains to form and maintain large transmission networks. The predominance of these strain networks, seen in two distant geographic regions of the former Soviet Union clearly limit the use of standardized MDR-TB therapies, e.g. the newly WHO endorsed short MDR-TB regimen, in these settings.

Temporal reconstruction of the resistance mutation acquisition and of changes in bacterial population sizes over three decades demonstrates that MDR outbreak strains already became resistant to both first- and second-line drugs in the 1980s. Fully first-line resistant strains massively expanded in the 1990s, a period that shortly preceded or immediately followed the end of the Soviet Union, years before the implementation of DOTS and programmatic second-line MDR-TB treatment. This is in line with the known rise in TB incidence that accompanied the economic breakdown in Russia during the 1990's (30).

From a bacterial genetic point of view, our data show that particular MDR and pre-XDR strain subgroups are highly transmissible despite accumulation of multiple resistance mutations. The acquisition of compensatory mutations after introduction of low fitness cost resistance mutations (e.g. *katG* S315T (10), *rpoB* S450L (8), *rpsL* K43R (31)) seems the critical stage allowing for higher transmission rates. Multiple regression analyses further strengthened this hypothesis by demonstrating that the presence of fitness compensating variants was positively associated with transmission success in different settings and outbreak clades, independently of the accumulation of resistance mutations. Compensatory evolution thus appears to play a central role in driving large MDR-TB epidemics such as that seen with the Beijing CAO-clade.

A particular concern is the high prevalence of mutations conferring resistance to second-line drugs currently included in treatment regimens, among the dominant MDR-MTBC strains. Their detected emergence in a period preceding DOTS implementation, e.g. in Karakalpakstan, can be explained by past, largely empirical treatment decisions or self-medication. For instance, high frequencies of mutations in the *ribD* promoter region, and *folC* among Beijing-CAO strains, associated with para-aminosalicylic acid resistance (32, 33), are a likely consequence of the use of para-aminosalicylic acid in failing treatment regimens in the late

1970s to the early 1980s in the Soviet Union (34–36). Likewise, the frequent independent emergence of mutations in the *eis* promoter and of rare variants in the upstream region of *whiB7*, both linked to resistance to aminoglycosides (mainly streptomycin and kanamycin) (37, 38), probably reflects self-administration of kanamycin that was available in local pharmacies.

The pre-existence of fully first-line resistant strain populations (e.g. CAO-Beijing in Karakalpakstan) likely contributed to the poor treatment outcomes observed among MDR-TB patients following the implementation of first-line DOTS treatment in 1998 (16). This period coincides with a detected CAO population size increase, likely reflecting the absence of drug susceptibility testing and therefore appropriate second-line treatment during extended hospitalization at the time, resulting in prolonged infectiousness of TB-patients and further spread of these strains.

The frequencies of fluoroquinolone resistance, mediated by *gyrA* and *gyrB* mutations, remained low among the Karakalpakstan MDR-MTBC strains, which is consistent with the notion that such drugs were rarely used for treating TB in former Soviet Union countries (see discussion (13), (34–36)). This observation explains the generally favorable MDR-TB treatment outcomes observed with the use of individualized second-line regimens, including a fluoroquinolone, in the latter MDR-TB treatment program in the Karakalpakstan patient population (14, 39). However, fluoroquinolone resistance, representing the last step towards XDR-TB, is already emerging as reported for strains in Beijing clade A and B (13).

In conclusion, the (pre-) existence and wide geographic dissemination of highly resistant and highly transmissible strain populations most likely contributes to increasing M/XDR-TB incidence rates despite scaling up of the MDR-TB programs in some Eastern European and Russian regions (15, 30, 40). Importantly, from the large spectrum of resistance detected among dominating strains in this study, it can be predicted that standardized therapies, including the newly WHO endorsed short MDR-TB regimen in Uzbekistan, are/will be largely ineffective for many patients in Samara and Karakalpakstan, and likely elsewhere in Eurasia. In order to successfully control the worldwide MDR-TB epidemics, universal access to rapid and comprehensive drug susceptibility testing, best supported by more advanced technologies, will

be crucial for guiding individualized treatment with existing and new/repurposed TB drugs and to maximize chances of cure and prevention of further resistance acquisition.

Funding

Parts of the work were funded by the German Ministry of Education and Research (BMBF) for the German Center of Infection Research (DZIF)

References and Notes:

1. WHO (2016) WHO treatment guidelines for drug-resistant tuberculosis - 2016 update. Available at: http://www.who.int/tb/MDRTBguidelines2016.pdf [Accessed September 20, 2016].

2. Middlebrook G, Cohn ML (1953) Some observations on the pathogenicity of isoniazid-resistant variants of tubercle bacilli. *Science* 118(3063):297–299.

3. Borrell S, Gagneux S (2009) Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis. *Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis* 13(12):1456–1466.

4. Billington OJ, McHugh TD, Gillespie SH (1999) Physiological cost of rifampin resistance induced in vitro in Mycobacterium tuberculosis. *Antimicrob Agents Chemother* 43(8):1866–1869.

5. Burgos M, DeRiemer K, Small PM, Hopewell PC, Daley CL (2003) Effect of drug resistance on the generation of secondary cases of tuberculosis. *J Infect Dis* 188(12):1878–1884.

6. Dye C, Espinal MA (2001) Will tuberculosis become resistant to all antibiotics? *Proc Biol Sci* 268(1462):45–52.

7. Andersson DI, Levin BR (1999) The biological cost of antibiotic resistance. *Curr Opin Microbiol* 2(5):489–493.

8. Gagneux S, et al. (2006) The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. *Science* 312(5782):1944–1946.

9. Müller B, Borrell S, Rose G, Gagneux S (2013) The heterogeneous evolution of multidrug-resistant Mycobacterium tuberculosis. *Trends Genet TIG* 29(3):160–169.

10. Pym AS, Saint-Joanis B, Cole ST (2002) Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. *Infect Immun* 70(9):4955–4960.

11. Comas I, et al. (2011) Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* 44(1):106–110.

12. Balabanova Y, et al. (2004) Antimicrobial prescribing patterns for respiratory diseases including tuberculosis in Russia: a possible role in drug resistance? *J Antimicrob Chemother* 54(3):673–679.

13. Casali N, et al. (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 46(3):279–286.

14. Cox HS, et al. (2007) Multidrug-resistant tuberculosis treatment outcomes in Karakalpakstan, Uzbekistan: treatment complexity and XDR-TB among treatment failures. *PloS One* 2(11):e1126.

15. Ulmasova DJ, et al. (2013) Multidrug-resistant tuberculosis in Uzbekistan: results of a nationwide survey, 2010 to 2011. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 18(42).

16. Cox H, et al. (2006) Tuberculosis recurrence and mortality after successful treatment: impact of drug resistance. *PLoS Med* 3(10):e384.

17. Kent P, Kubica G (1985) *Public Health Mycobacteriology: A guide for the level III laboratory.* (US Department of Health and Human Services, Centres for Disease Control, Atlanta, Ga).

18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* 25(14):1754–1760.

19. McKenna A, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.

20. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl* 25(16):2078–2079.

21. Ford CB, et al. (2011) Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* 43(5):482–486.

22. Walker TM, et al. (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 13(2):137–146.

23. Roetzer A, et al. (2013) Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10(2):e1001387.

24. Walker TM, et al. (2014) Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2(4):285–292.

25. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 5(3):e9490.

26. Rambaut A, Lam TT, Max Carvalho L, Pybus OG (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2(1). doi:10.1093/ve/vew007.

27. Bouckaert R, et al. (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Comput Biol* 10(4):e1003537.

28. Casali N, et al. (2012) Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res* 22(4):735–745.

29. Merker M, et al. (2015) Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat Genet* 47(3):242–249.

30. Institute of Medicine (US) Forum on Drug Discovery D, Science RA of M (2011) Drug-Resistant Tuberculosis in the Russian Federation. Available at: http://www.ncbi.nlm.nih.gov/books/NBK62453/ [Accessed October 12, 2015].

31. Böttger EC, Springer B, Pletschette M, Sander P (1998) Fitness of antibiotic-resistant microorganisms and compensatory mutations. *Nat Med* 4(12):1343–1344.

32. Zheng J, et al. (2013) para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. *J Biol Chem* 288(32):23447–23456.

33. Zhao F, et al. (2014) Binding pocket alterations in dihydrofolate synthase confer resistance to para-aminosalicylic acid in clinical isolates of Mycobacterium tuberculosis. *Antimicrob Agents Chemother* 58(3):1479–1487.

34. Ministry of Health of the USSR (1963) Methodological recommendations. *Chemotherapy in TB Treatment* (Moscow).

35. Ministry of Health of the USSR (1983) Methodological recommendations. *Chemotherapy in TB Treatment* (Moscow).

36. Mishin VY (2008) TB chemotherapy (review). 34–43.

37. Zaunbrecher MA, Sikes RD Jr, Metchock B, Shinnick TM, Posey JE (2009) Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 106(47):20004–20009.

38. Reeves AZ, et al. (2013) Aminoglycoside cross-resistance in Mycobacterium tuberculosis due to mutations in the 5' untranslated region of whiB7. *Antimicrob Agents Chemother* 57(4):1857–1865.

39. Lalor M, et al. (2011) Treatment outcomes in multidrug-resistant TB pateints in Uzbekistan (conference abstract) (Paris, France).

40. Medecins Sans Frontieres (2013) *International Activity Report* (Medecins Sans Frontieres).

Acknowledgments

We thank: I. Razio, P. Vock, T. Ubben and J. Zallet from Borstel, Germany for technical assistance; the national and expatriate staff of Médecins Sans Frontières, Karakalpakstan; Dr. Atadjan and Dr. K. Khamraev from the Ministry of Health (Karakalpakstan) for their support. Parts of this work have been supported by the European Union TB-PAN-NET (FP7-223681) project and the German Center for Infection Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Raw sequence data (fastq files) have been deposited at the European Nucleotide Archive (ENA) under the project number (pending).

Author contributions

S.N., M.M., T.W. H.C. and P.S. designed the study. M.M., M.B., H.C., S.F., J.P.R., U.N., R.D., S.B., S.G., V.N., S.A., and T.W. analyzed data e.g. classical mycobacteriology, performed population genetic, phylogenetic and statistical analysis. T.A.K. performed whole genome sequencing and variant calling. All authors analyzed the data and contributed to data interpretation and manuscript writing.

Conflicts of interest:

None to declare.

Figures

Fig. 1: Drug resistance and transmission success among MDR-MTBC strains from Karakalpakstan, Uzbekistan. Maximum likelihood phylogeny (GTR substitution model, 1,000 resamples) of 277 MDR-MTBC strains from Karakalpakstan, Uzbekistan sampled from 2001 to 2006. Columns show drug resistance associated mutations to first- and second-line drugs (different mutations represented by different colors), genetic classification of pre-XDR (purple) and XDR (pink) strains, and putative compensatory mutations in the RNA polymerase genes *rpoA*, *rpoB* and *rpoC*. Transmission index represents number of isolates within a maximum range of 10 SNPs at whole genome level. MTBC-Beijing strains (lineage 2) are differentiated into three sub-groups (i.e. Central Asian Outbreak (CAO), group 2 and group 3). Strains belonging to lineage 4 (Euro-American) are colored in grey: H=isoniazid, R=rifampicin, S=streptomycin, E=ethambutol, Z=pyrazinamide, FQ=fluoroquinolone, AG=aminoglycosides, Km=kanamycin Cm=capreomycin, TA=thioamide, PAS=para-aminosalicylic acid.

Fig. 2: Evolutionary history of MTBC Central Asian outbreak (CAO) strains

Genealogical tree of CAO strains in Karakalpakstan, Uzbekistan and effective population size over time based on a (piecewise-constant) Bayesian skyline approach using the GTR substitution model and a strict molecular clock prior of 0.94×10^{-7} substitutions per nucleotide per year. Pink shaded area represents changes in the effective population size giving the 95% highest posterior density (HPD) interval with the pink line representing the mean value. Vertical lines indicate time points of the implementation of the first standardized TB treatment program (DOTS) in Karakalpakstan and of the declaration of Uzbekistan as independent republic. Symbols on branches show steps of fixation of resistance conferring mutations.

Fig. 3: Compensatory mutations and drug resistance levels

Comparisons between strains carrying compensatory mutations (in orange) and strains with no-compensatory mutations (in blue), from the Karakalpakstan dataset. A) Boxplot showing number of

resistance mutations for the two categories (without or with compensatory mutations). The two categories were significantly different (two-sample t-test P=1.2x10^{-10}). B) Bubble plots showing the transmission index (number of strains differing by less than 10 SNPs) as a function of antibiotic resistance related mutations. Bubble sizes are proportional to the numbers of strains. C) Density plot of the number of resistance-conferring mutations for 4 groups of strains sourced from the Karakalpakstan data. Proportions are adjusted by using Gaussian smoothing kernels. The 4 groups are composed of non-CAO strains with no compensatory mutations; non-CAO strains carrying compensatory mutations; CAO strains with no compensatory mutations and CAO strains carrying compensatory mutations. These groups are respectively colored in light blue, dark blue, light orange and light red.

Fig. 4: Contributions of resistance-conferring and compensatory mutations to the transmission success of *M. tuberculosis* of the Beijing-CAO clade, Karakalpakstan, Uzbekistan. Shown are the coefficients and 95% confidence bands of multiple linear regression of the transmission success score, defined as the size of clusters diverging by at most *N* SNPs and divided by *N* or, equivalently, the size of clusters that evolved over *N* years divided by *N*. The presence of compensatory mutations was independently associated with transmission success, with a maximum association strength found for SNP distances ranging from 10 to 20 SNPs, corresponding to transmission clusters beginning around 1995.

Fig. 5: MDR-MTBC phylogeny and resistance mutations of strains from Samara (Russia) and Karakalpakstan (Uzbekistan)

Maximum likelihood tree (with 1,000 resamples, GTR nucleotide substitution model) based on 12,567 variable positions (SNPs) among 705 MDR-MTBC isolates from Karakalpakstan and Samara. Any resistance associated mutations (see methods) for individual antibiotics are depicted with red bars for each strain. The presence of any putative compensatory mutation in the RNA polymerase genes *rpoA*, *rpoB*, *rpoC* is depicted with green bars and country of origin and a genotypic pre-XDR and XDR strain classification is color coded. PAS = para-aminosalycylic acid.

Fig. 6: Percentage of drug resistance among 705 MDR-MTBC strains from Samara (Russia) and Karakalpakstan (Uzbekistan)

MDR-MTBC strains stratified to three Beijing sub-groups, other Beijing strains and non-Beijing strains. Proportions of strains with identified molecular drug resistance mutations (see additional data) which mediate resistance to multiple first- and second-line anti-TB drugs. Values are rounded. Drugs used in the WHO endorsed standardized short MDR-TB regimen marked with grey boxes.

*The short MDR-TB regimen further includes high-dose isoniazid treatment, and clofazimine. In that regard, we identified 622/705 (85.4%) of the MDR-MTBC strains with the well-known high-level isoniazid resistance mediating mutation *katG* S315T (additional data), for clofazimine resistance mediating mutations are not well described.

Supplementary Figures and Tables

Fig. S1: Box-Plot showing pairwise SNP distances among identified Beijing genotype subgroups in comparison to non-Beijing strains. Box represents inter quartile range, whiskers represent 95% of the data, outliers shown as black dots; solid black line represents the median.

Fig. S2: Proportions of different genome-based subgroups in Karakalpakstan, Uzbekistan stratified to the years 2001/02, 2003/04, 2005/06.

Fig. S3: Box-Plot showing median number of (A) phenotypic and (B) genotypic drug resistances (in addition to the MDR classification, i.e. isoniazid and rifampicin resistance). Box represents inter quartile range, whiskers represent 95% of the data, outliers shown as black dots; solid black line represents the median. Beijing CAO strains exhibit more phenotypic drug resistances compared to non-Beijing strains

529 (P=0.0079) and more genotypic drug resistances compared to other Beijing strains (P<0.0001), and non-
530 Beijing strains (P<0.0001).
531
532 **Fig. S4:** Linear regression analysis showing correlation between root-to-tip distance and sampling years of
533 an extended collection of 220 Beijing CAO datasets covering the period 1995 to 2009.
534
535 **Fig S5:** Number of drug resistance mutations among different MDR-MTBC groups from Samara (n=428)
536 and Karakapakstan (n=277). Box-Plot with mean (diamond) and median (horizontal line) number of
537 genotypic drug resistances (see methods) to additional anti-TB drugs (beyond MDR defining rifampicin
538 and isoniazid resistance). Box represents inter quartile range, whiskers represent 95% of the data, outliers
539 shown as black dots. P-values for three major Beijing outbreak clades (A, B and CAO), and non-Beijing
540 strains (mainly lineage 4 isolates) were calculated with unpaired t-tests with Welch correction compared
541 to the group 'other Beijing' strains. Color codes according to Fig. 5. P-values. Clade A (P≤0.0001), Clade
542 B (P=0.0143), CAO (P≤0.0001), and non-Beijing (P=0.0009).
543
544 **Fig. S6:** Comparisons between strains carrying compensatory mutations (in orange) and strains with no-
545 compensatory mutations (in blue), from the Samara dataset. A) Boxplot showing number of resistance
546 mutations for the two categories (without or with compensatory mutations). The two categories were
547 significantly different (two-sample t-test $P<2.2\times10^{-16}$). B) Bubble plots showing the transmission index
548 (number of strains differing by less than 10 SNPs) as a function of antibiotic resistance related mutations.
549 Bubble sizes are function of the number of strains. C) Density plot of the number of resistance-conferring
550 mutations for strains carrying compensatory mutations (orange) and strains that don't carry compensatory
551 mutation (blue) from Samara dataset. Proportions are adjusted by using Gaussian smoothing kernels.
552
553 **FigS7:** Contributions of resistance-conferring and compensatory mutations to the transmission success of
554 *M. tuberculosis* of the Beijing-A clade from Samara, Russia. Shown are the coefficients and 95%

555 confidence bands of multiple linear regression of the transmission success score, defined as the size of
556 clusters diverging by at most *N* SNPs and divided by *N* or, equivalently, the size of clusters that evolved
557 over *N* years divided by *N*. Compensatory mutations were independently associated with transmission
558 success, with a maximum association strength found for transmission clusters beginning around 1999.
559
560 **Table S1**: Main characteristics of patients from cohorts 1 and 2 from in Karakalpakstan.

	Cohort 1	**Cohort 2**	**P value**
Year of strain collection (patient diagnosis with MDR-TB)	2001-2002	2003-2006	
No. MDR cases diagnosed within time period	57	300	
No. Included in this analysis	49 (86%)	228 (76%)	0.094
Reasons for non-inclusion:			
Multiple strain infection	6	1	
No DNA available	2	40	
Patient already in cohort 1	NA	11	
MIRU not available	0	20	
Patient residence (within Karakalpakstan)			
Nukus	34 (69%)	146 (64%)	0.49
Chimbay	6	64	
Other	9	1	
Unknown	0	17	
Male	27 (55%)	119 (52%)	0.72
Age (median, IQR)	32, 27-38	31, 24-41	0.40
Missing age	0	49 (21%)	
Previous TB treatment	38 (78%)	228 (100%)	<0.0001
First-line resistance profile:			
HR	1	2	
HRE	0	1	
HRS	12 (24%)	41 (18%)	
HRES	28 (57%)	49 (21%)	
HRSZ	1 (2%)	27 (12%)	
HREZ	1	1	
HRESZ	7 (14%)	107 (47%)	<0.0001
No. of first-line drugs resistant			
2	1	2	
3	12 (24%)	42 (18%)	
4	30 (61%)	77 (34%)	
5	7 (14%)	107 (47%)	<0.0001
Availability of second-line drug susceptibility testing (DST)	Ofx, Cap, Proth	Ofx, Cap, Ami, Eth, Cyc, PAS	
Ofx resistance	5 (10%)	6 (3%)	0.033
Cap resistance	1 (2%)	53 (23%)	0.0001

561

562 Abbreviations: H=isoniziad, R=rifampicin, E=ethambutol, S=streptomycin, Z=pyrazinamide,
563 Ofx=ofloxacin, Cap=capreomycin, Proth=Prothionamide, Ami=Amikacin, Eth=ethionamide,
564 Cyc=cycloserine, PAS=para-aminosalicylic acid

565

566 **Table S2**

567 Likelihood scores for different substitution models calculated with Jmodeltest 2.1 and statistical model
568 selection based on Akaike and Bayesian Information Criteration (AIC and BIC). Best model is assumed to
569 have the lowest criteration value. Shown are the top 10 AIC models. Substitution models used for Bayesian
570 inference marked in bold.

571

Subst. model	-lnL	AIC	Δ AIC (AIC ranking)	BIC	Δ BIC (BIC ranking)
GTR	8837.6437	**18567.2875**	**0.0 (1)**	**21041.0025**	**7.0748 (2)**
GTR+I	8837.6747	18569.3494	2.0619 (2)	21048.6109	14.6832 (5)
GTR+G	8838.9842	18571.9684	4.6809 (3)	21051.2299	17.3022 (6)
GTR+I+G	8839.0077	18574.0153	6.7278 (4)	21058.8233	24.8955 (8)
TPM1uf	8845.426	18576.852	9.5645 (5)	21033.9277	0.0 (1)
TPM1uf+I	8845.4446	18578.8891	11.6016 (6)	21041.5113	7.5836 (3)
TPM1uf+G	8846.7354	18581.4709	14.1834 (7)	21044.093	10.1653 (4)
TPM1uf+I+G	8846.7697	18583.5395	16.252 (8)	21051.7081	17.7804 (7)
SYM	8860.6478	18607.2955	40.008 (9)	21064.3712	30.4435 (9)
SYM+I	8860.6826	18609.3652	42.0777 (10)	21071.9874	38.0596 (12)

572

573

574

575

576

577

578

579

Table S3:

Path sampling results and model selection based on Δ marginal L estimates (relative to best model, given in bold fonts) considering 75 path sampling steps and chain lengths of 15 million analysing Beast runs of a combined dataset of Central Asian outbreak (CAO) isolates originated from Germany (1995-2000), Karakalpakstan (2001-2006), and Samara (2008-2010).

Subst. model	Clock model	Demographic model	Marginal L estimate	Mean ESS	Δ marginal L estimate	Subst rate x 10^{-7} (95%HPD)
GTR	Strict (no tip dating)	Coalescent constant size	-10131.67	214.3	32.21	1.0 (fixed)
GTR	**Strict (tip dating)**	**Coalescent constant size**	**-10099.46**	**153.7**	**ref**	**1.0 (fixed)**
GTR	Relaxed, lognormal	Coalescent constant size	-10117.21	123.9	17.75	0.96 (0.65-1.24)
Strict clock with tip dating selected as best model						
GTR	Relaxed, lognormal	Coalescent constant size	-10117.21	123.9	78.28	0.96 (0.65-1.24)
GTR	Relaxed, lognormal	Exponential	-10044.41	189.5	5.48	0.88 (0.58-1.21)
GTR	**Relaxed, lognormal**	**Bayesian skyline**	**-10038.93**	**98.6**	**ref**	**0.94 (0.72-1.15)**
Mutation rate of 0.94 used for a Bayesian skyline model with a strict molecular clock for CAO strains from Karakalpakstan only						

Abbreviations: HPD=Highest posterior density interval, GTR=generalized time reversible

Table S4: Mutations in *rpoB*, *rpoA* and *rpoC* associated with a putative compensatory effect in rifampicin resistant MTBC strains. Data from 277 MDR-MTBC strains from Karakalpakstan, Uzbekistan, stratified to the particularly successful variant termed Central Asian outbreak (CAO) and other Beijing strains. Pairwise differences between the two groups calculated with Fisher exact test; two-tailed *P*-values are reported.

	Beijing CAO (n=173)	Other Beijing (n=64)	*P*-value	All (n=277)
rpoB mutations outside RRDR, excluding codon 170,400,491 variants	25 (14.5%)	12 (18.8%)		43 (15.5%)
wild type	147 (85.0%)	52 (81.3%)	0.43	234 (84.5%)
rpoC variants	95 (54.9%)	18 (28.1%)	0.0002	126 (45.5%)
wild type	78 (45.1%)	46 (71.2%)		151 (54.5%)
rpoA variants	5 (2.9%)	2 (3.1%)	1.00	7 (2.5%)
wild type	168 (97.1%)	62 (96.9%)		270 (97.5%)

Abbreviations: CAO=Central Asian outbreak, RRDR=rifampicin resistance determining region

Table S5: Proportions of genotypic drug resistance rates for different anti-TB drugs (beyond isoniazid and rifampicin resistance) and pre-XDR/XDR-TB classification among 705 MDR-MTBC clinical isolates from Samara (n=428) and Karakalpakstan (n=277), stratified to three identified major phylogenetic groups within the Beijing genotype/lineage and to other Beijing strains, and to non-Beijing strains (mainly lineage 4, Euro-American).

group	S	E	Z	Km	Am	Cm	Fq	Thio	PAS	Pre-XDR XDR
Beijing CAO (n=201)	201/201 100.0%	195/201 97.0%	152/201 75.6%	97/201 48.3%	37/201 18.4%	37/201 18.4%	6/201 3.0%	121/201 60.2%	99/201 49.3%	100/201 49.8%
Beijing clade B (W148) (n=103)	103/103 100.0%	83/103 80.6%	44/103 42.7%	61/103 59.2%	18/103 17.5%	18/103 17.5%	23/103 22.3%	75/103 72.8%	12/103 11.7%	64/103 62.1%
Beijing clade A (n=187)	184/187 98.4%	183/187 97.9%	163/187 87.2%	177/187 94.7%	0/187 0.0%	0/187 0.0%	33/187 17.6%	180/187 96.3%	7/187 3.7%	179/187 95.7%
Other Beijing (n=100)	91/100 91.0%	73/100 73.0%	52/100 52.0%	39/100 39.0%	20/100 20.0%	23/100 23.0%	14/100 14.0%	32/100 32.0%	15/100 15.0%	45/187 24.1%
Non-Beijing (n=114)	69/114 60.5%	63/114 55.3%	30/114 26.3%	39/114 34.2%	14/114 12.3%	14/114 12.3%	3/114 2.6%	34/114 29.8%	34/114 29.8%	40/114 35.1%

Abbreviations: S=Streptomycin, E=Ethambutol, Z=Pyrazinamide, Km=Kanamycin, Ami=Amikacin, Cap=Capreomycin, Fq=Fluoroquinolone, Thio=Thioamide, PAS=Para-aminosalicylic acid

Additional data:

34 Resistance targets and considered molecular markers.

Phylogenetic variants in 34 resistance associated target genes.

DST phenotypes and polymorphisms in resistance and compensatory genes found in 705 MDR-MTBC strains from Karakalpakstan, Uzbekistan and Samara, Russia and 19 CAO-Beijing strains from Germany.

Genotypic classification, transmission indexes, Accession numbers.

38 Central Asian outbreak (CAO) specific SNPs with annotations.

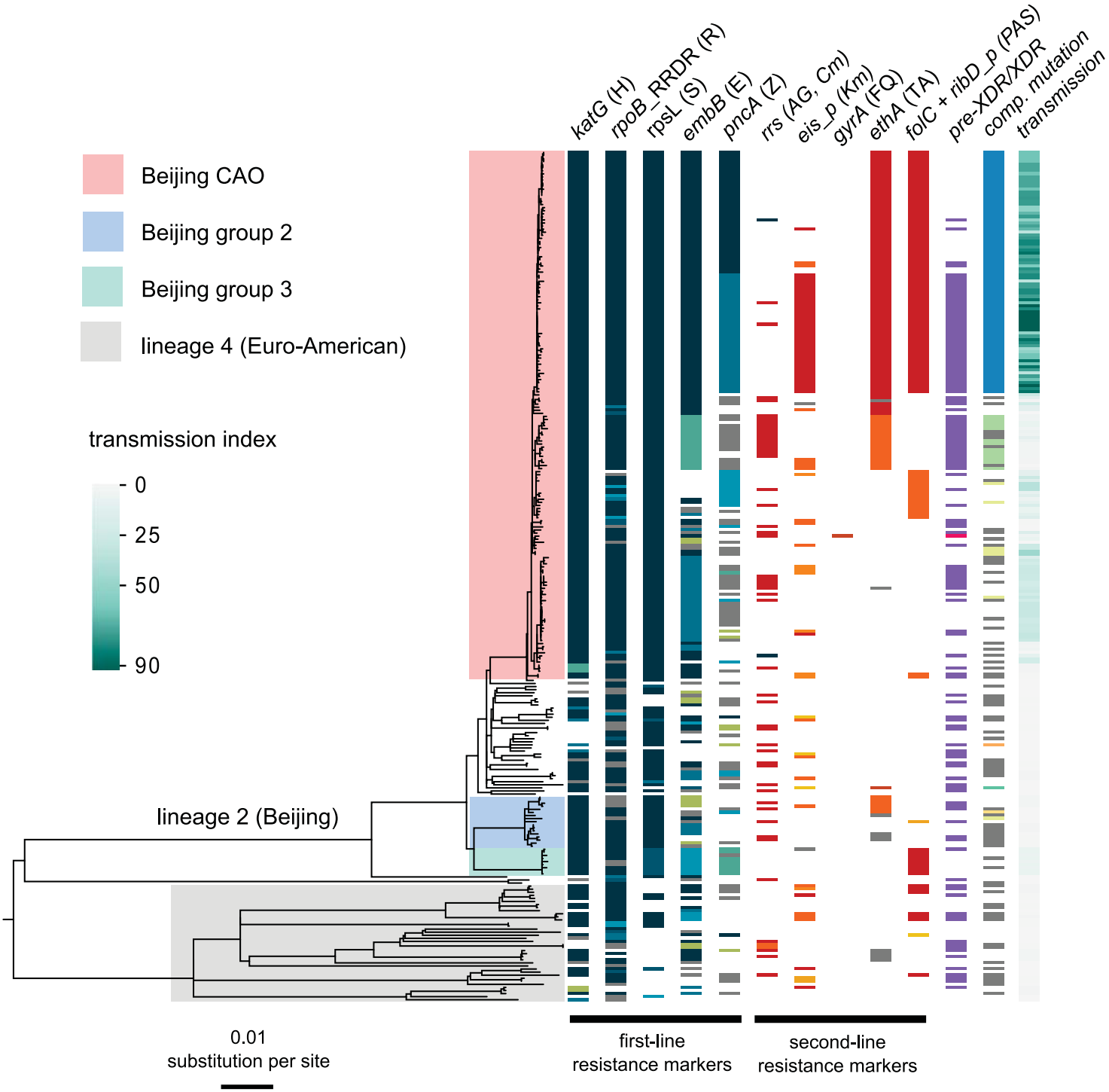

katG (H)
rpoB_RRDR (R)
rpsL (S)
embB (E)
pncA (Z)
rrs (AG, Cm)
eis_p (Km)
gyrA (FQ)
ethA (TA)
folC + ribD_p (PAS)
pre-XDR/XDR
comp. mutation
transmission
Beijing CAO
Beijing group 2
Beijing group 3
lineage 4 (Euro-American)
transmission index
0
25
50
90
lineage 2 (Beijing)
0.01
substitution per site
first-line
resistance markers
second-line
resistance markers

Figure 2

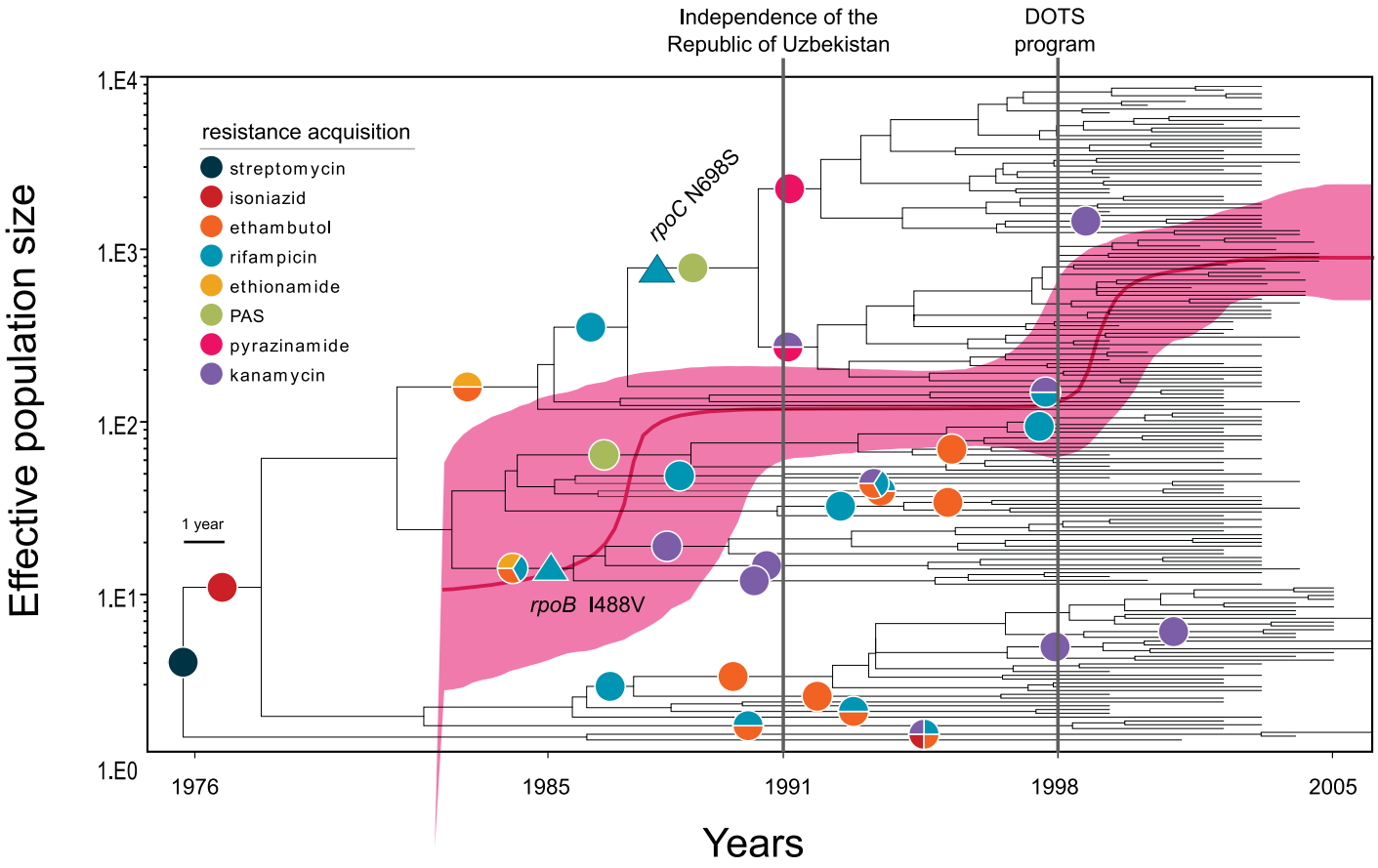

Independence of the Republic of Uzbekistan
DOTS program
resistance acquisition
streptomycin
isoniazid
ethambutol
rifampicin
ethionamide
PAS
pyrazinamide
kanamycin
rpoC N698S
rpoB I488V
1 year
Effective population size
1.E4
1.E3
1.E2
1.E1
1.E0
1976
1985
1991
1998
2005
Years

Figure 3

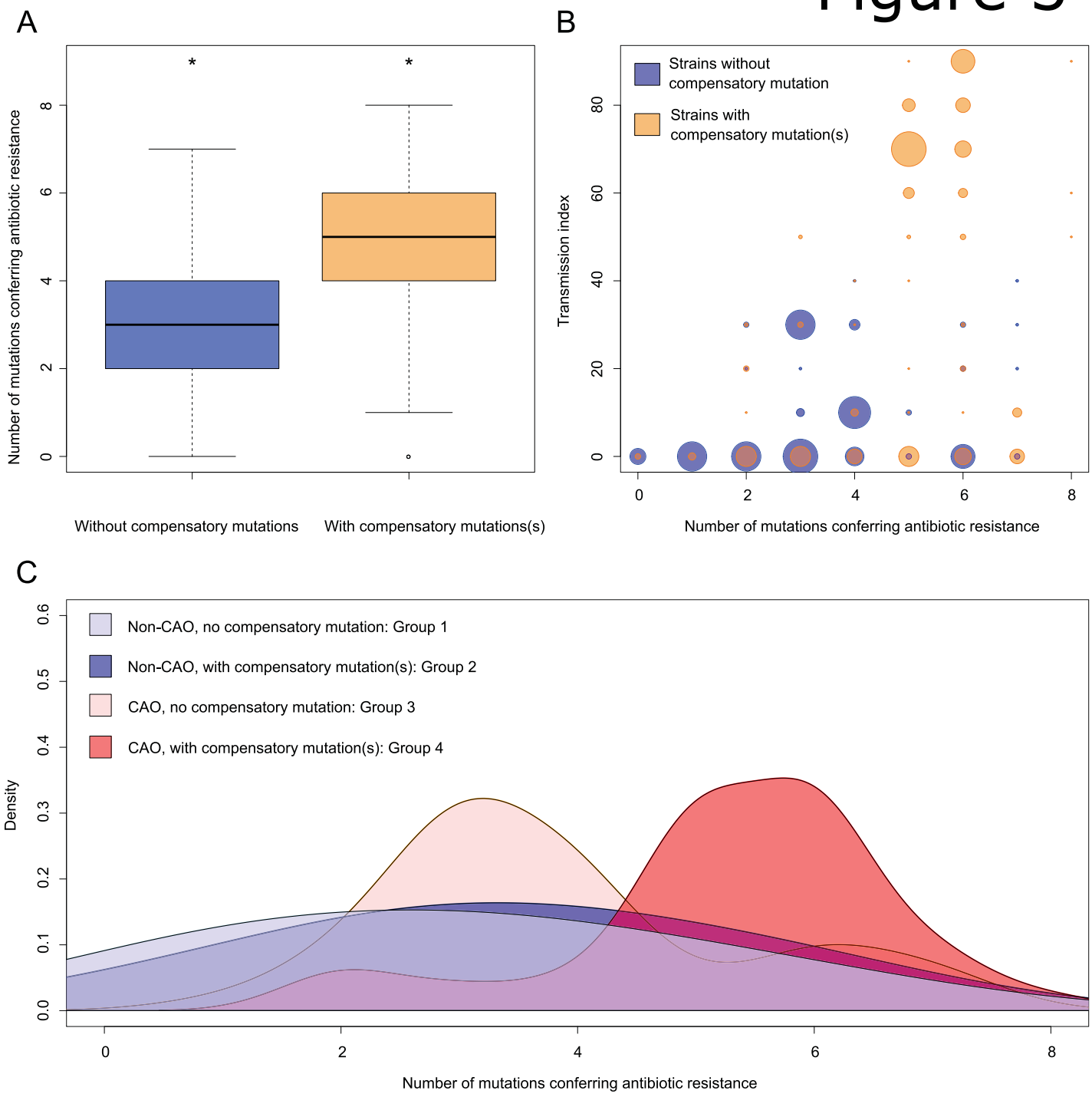
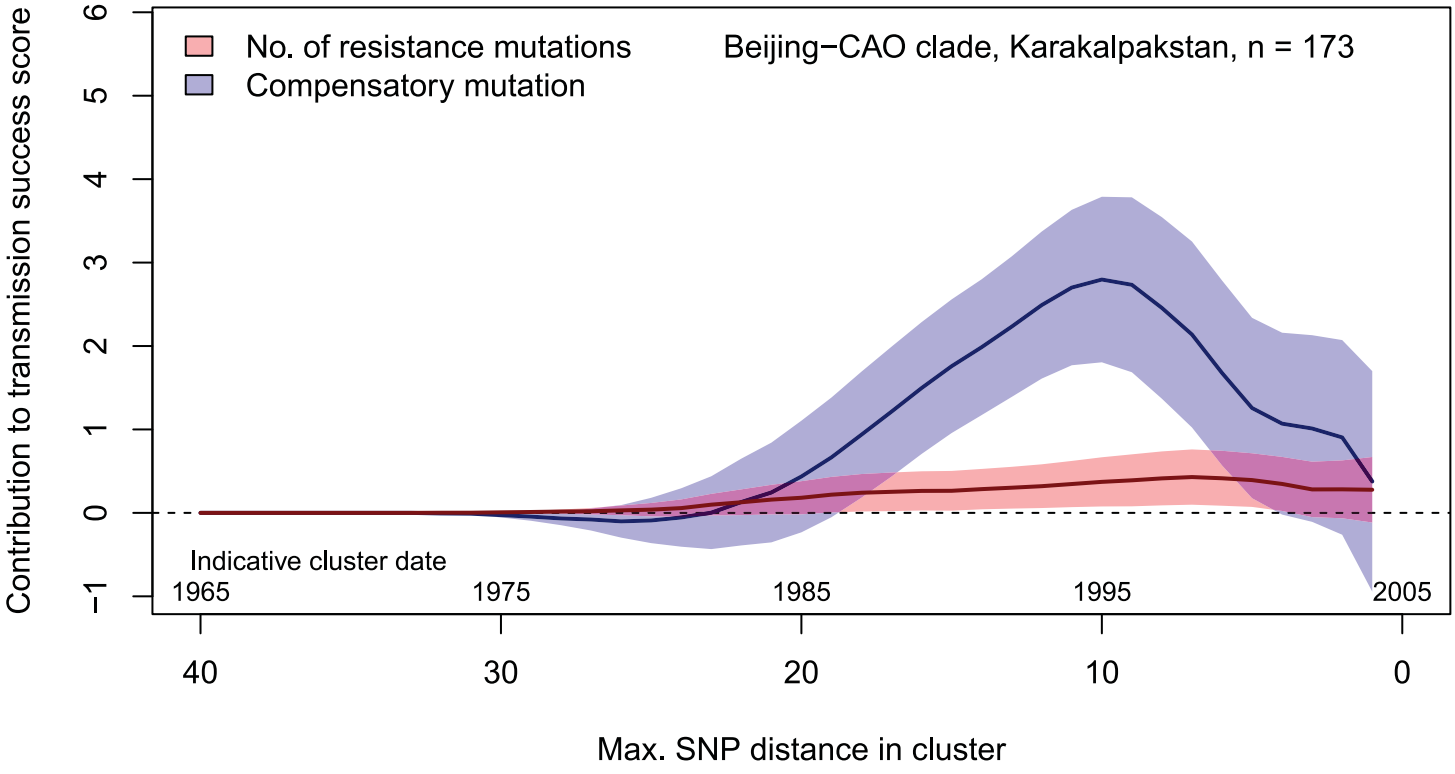

A
Number of mutations conferring antibiotic resistance
0
2
4
6
8
Without compensatory mutations
With compensatory mutations(s)
B
Transmission index
0
20
40
60
80
Strains without compensatory mutation
Strains with compensatory mutation(s)
0
2
4
6
8
Number of mutations conferring antibiotic resistance
C
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Non-CAO, no compensatory mutation: Group 1
Non-CAO, with compensatory mutation(s): Group 2
CAO, no compensatory mutation: Group 3
CAO, with compensatory mutation(s): Group 4
0
2
4
6
8
Number of mutations conferring antibiotic resistance

Figure 4


No. of resistance mutations
Compensatory mutation
Beijing−CAO clade, Karakalpakstan, n = 173
Contribution to transmission success score
−1
0
1
2
3
4
5
6
Indicative cluster date
1965
1975
1985
1995
2005
40
30
20
10
0
Max. SNP distance in cluster

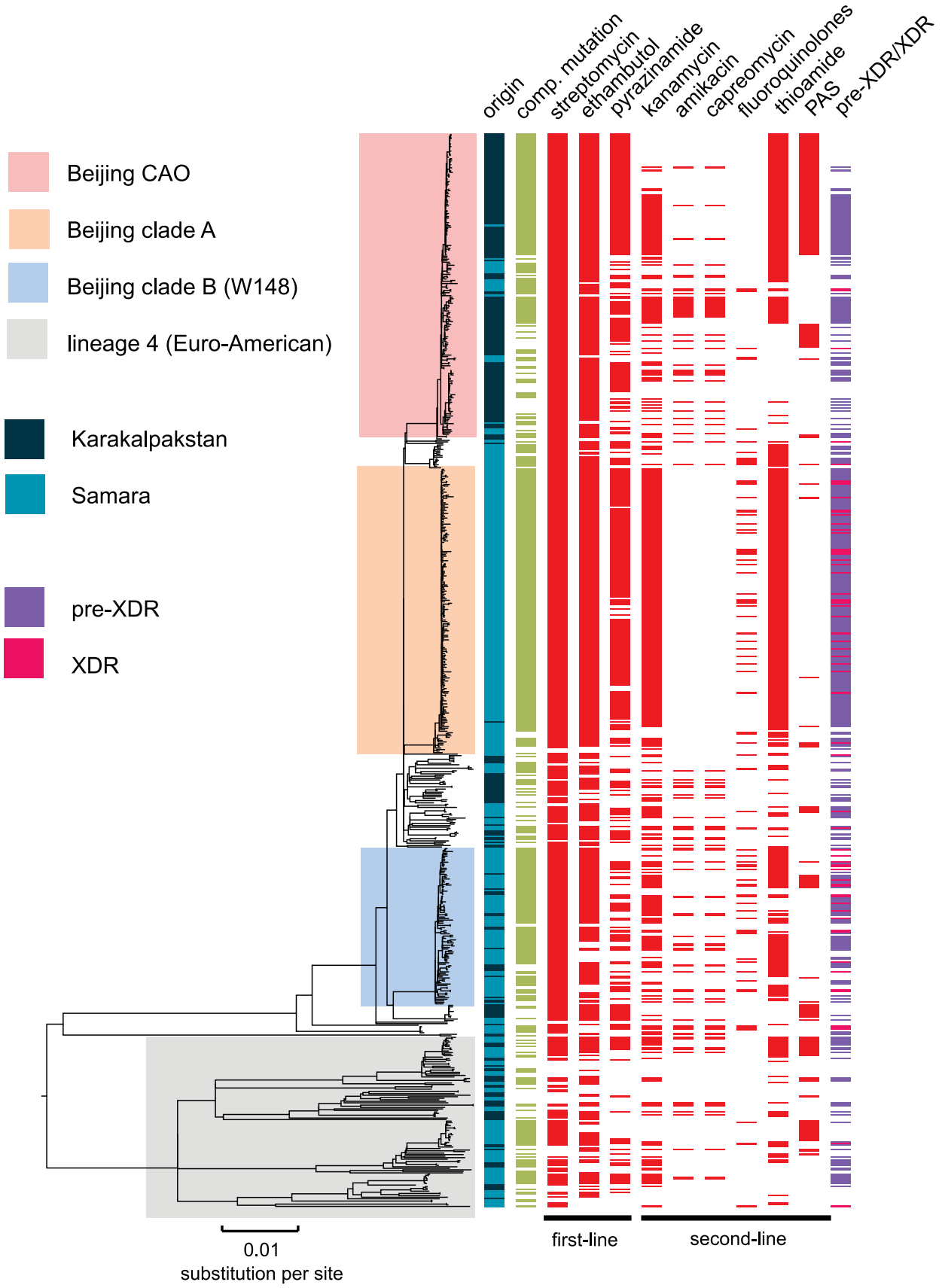

origin
comp. mutation
streptomycin
ethambutol
pyrazinamide
kanamycin
amikacin
capreomycin
fluoroquinolones
thioamide
PAS
pre-XDR/XDR
Beijing CAO
Beijing clade A
Beijing clade B (W148)
lineage 4 (Euro-American)
Karakalpakstan
Samara
pre-XDR
XDR
first-line
second-line
0.01
substitution per site

Figure 6

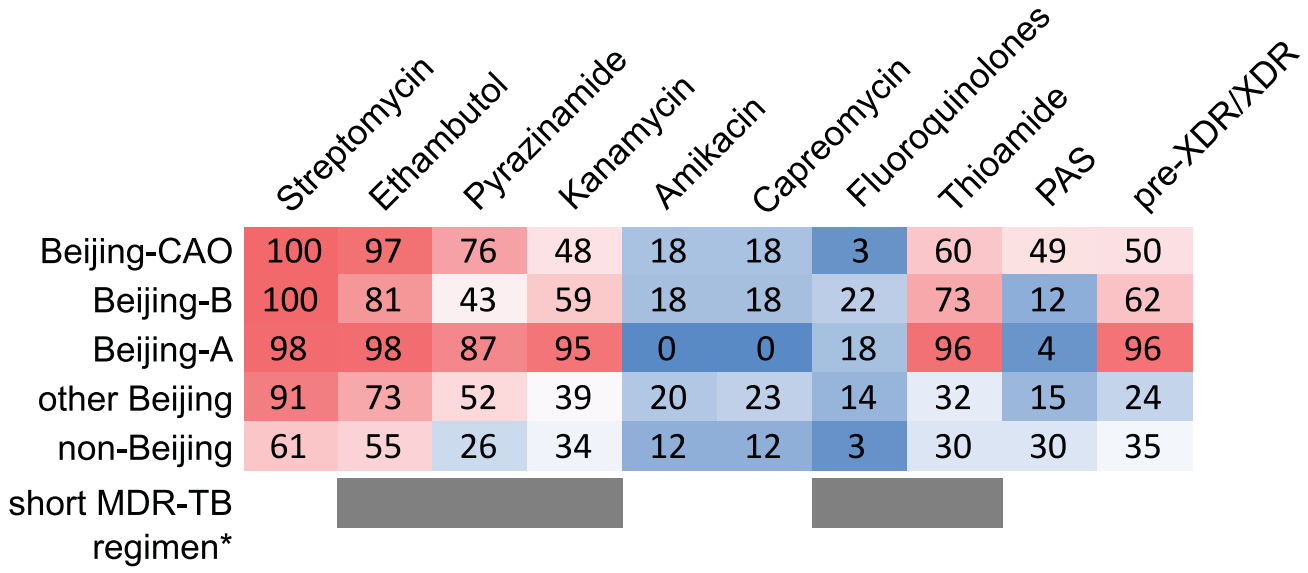

Streptomycin
Ethambutol
Pyrazinamide
Kanamycin
Amikacin
Capreomycin
Fluoroquinolones
Thioamide
PAS
pre-XDR/XDR
Beijing-CAO 100 97 76 48 18 18 3 60 49 50
Beijing-B 100 81 43 59 18 18 22 73 12 62
Beijing-A 98 98 87 95 0 0 18 96 4 96
other Beijing 91 73 52 39 20 23 14 32 15 24
non-Beijing 61 55 26 34 12 12 3 30 30 35
short MDR-TB regimen*

Fig.S1

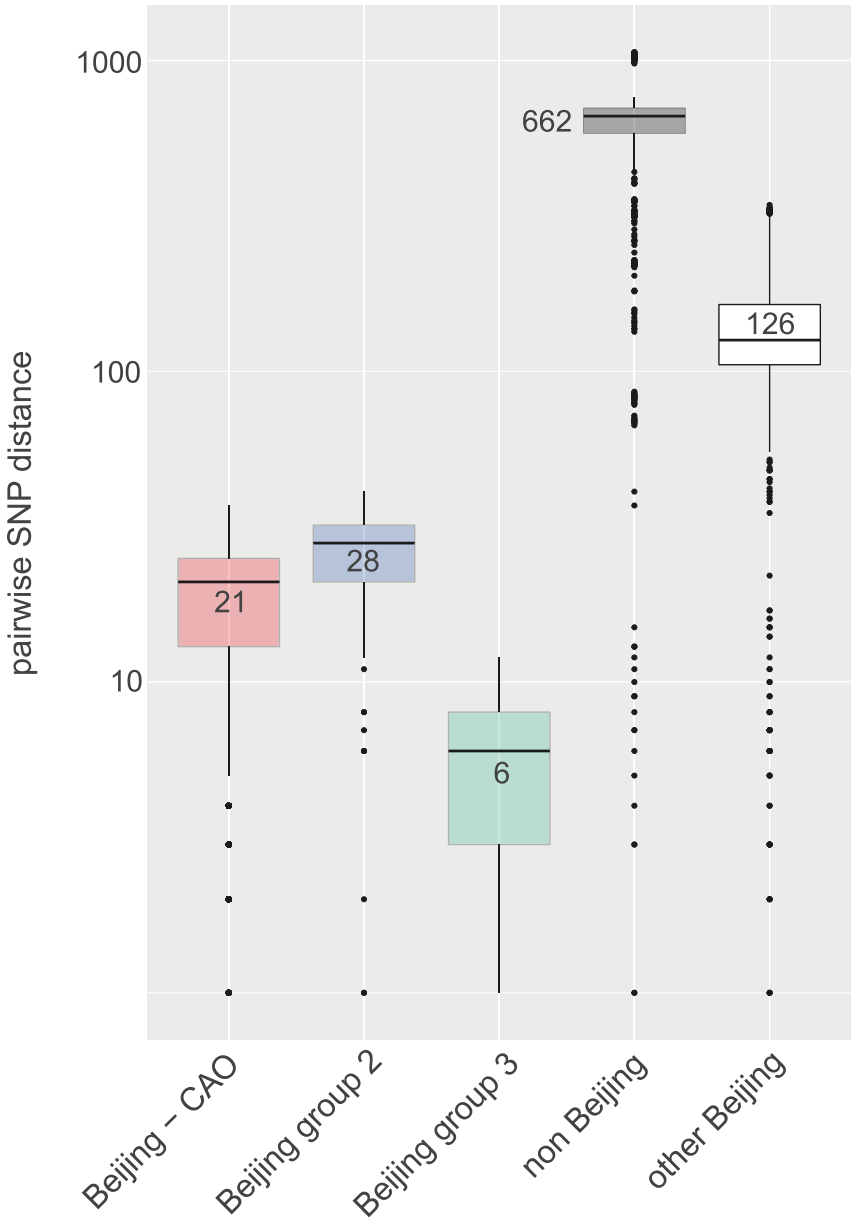

1000
100
10
pairwise SNP distance
21
28
6
662
126
Beijing – CAO
Beijing group 2
Beijing group 3
non Beijing
other Beijing

Fig.S2

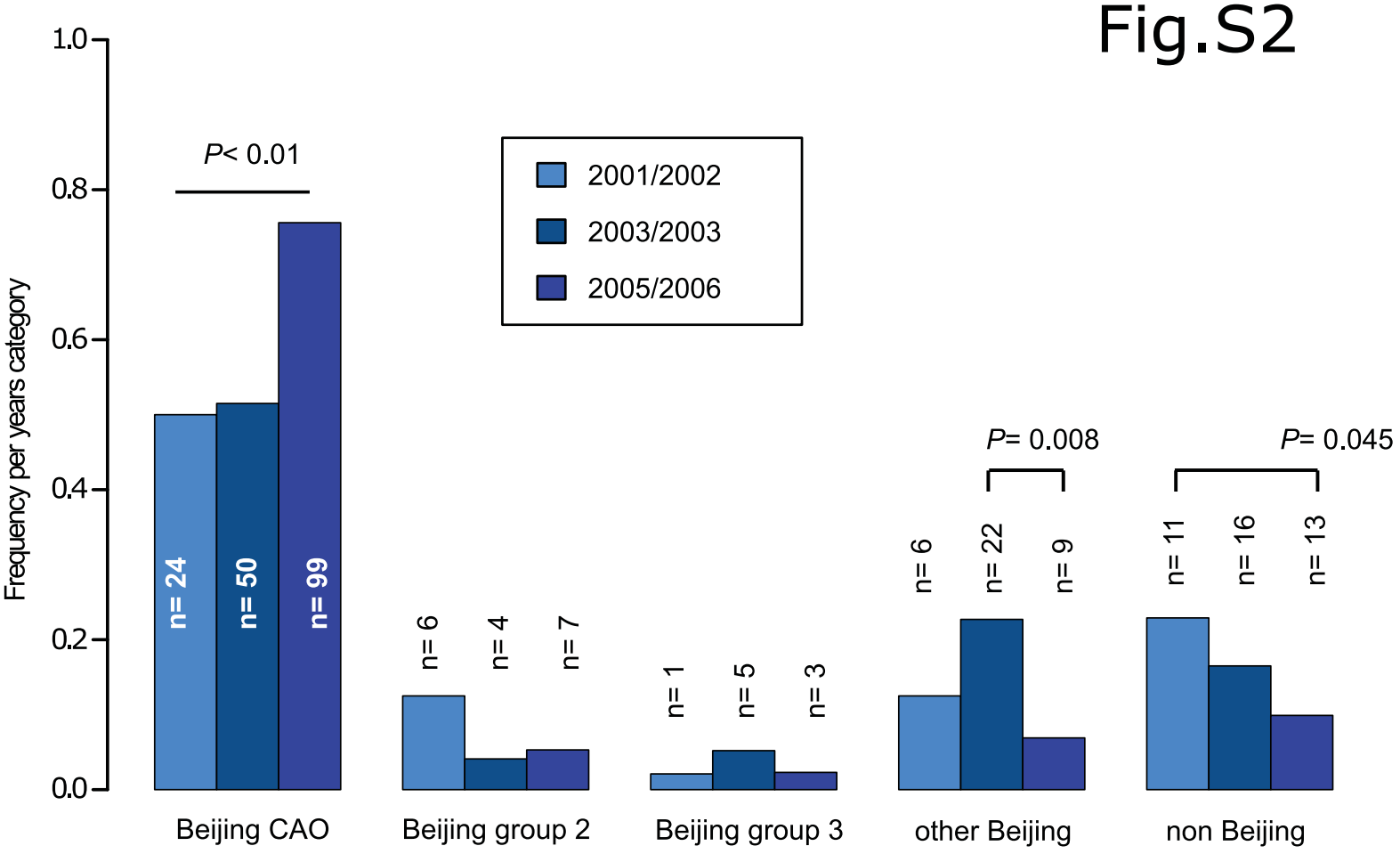

1.0
0.8
0.6
0.4
0.2
0.0
Frequency per years category
P< 0.01
2001/2002
2003/2003
2005/2006
n= 24
n= 50
n= 99
n= 6
n= 4
n= 7
n= 1
n= 5
n= 3
P= 0.008
n= 6
n= 22
n= 9
P= 0.045
n= 11
n= 16
n= 13
Beijing CAO
Beijing group 2
Beijing group 3
other Beijing
non Beijing

Fig.S3

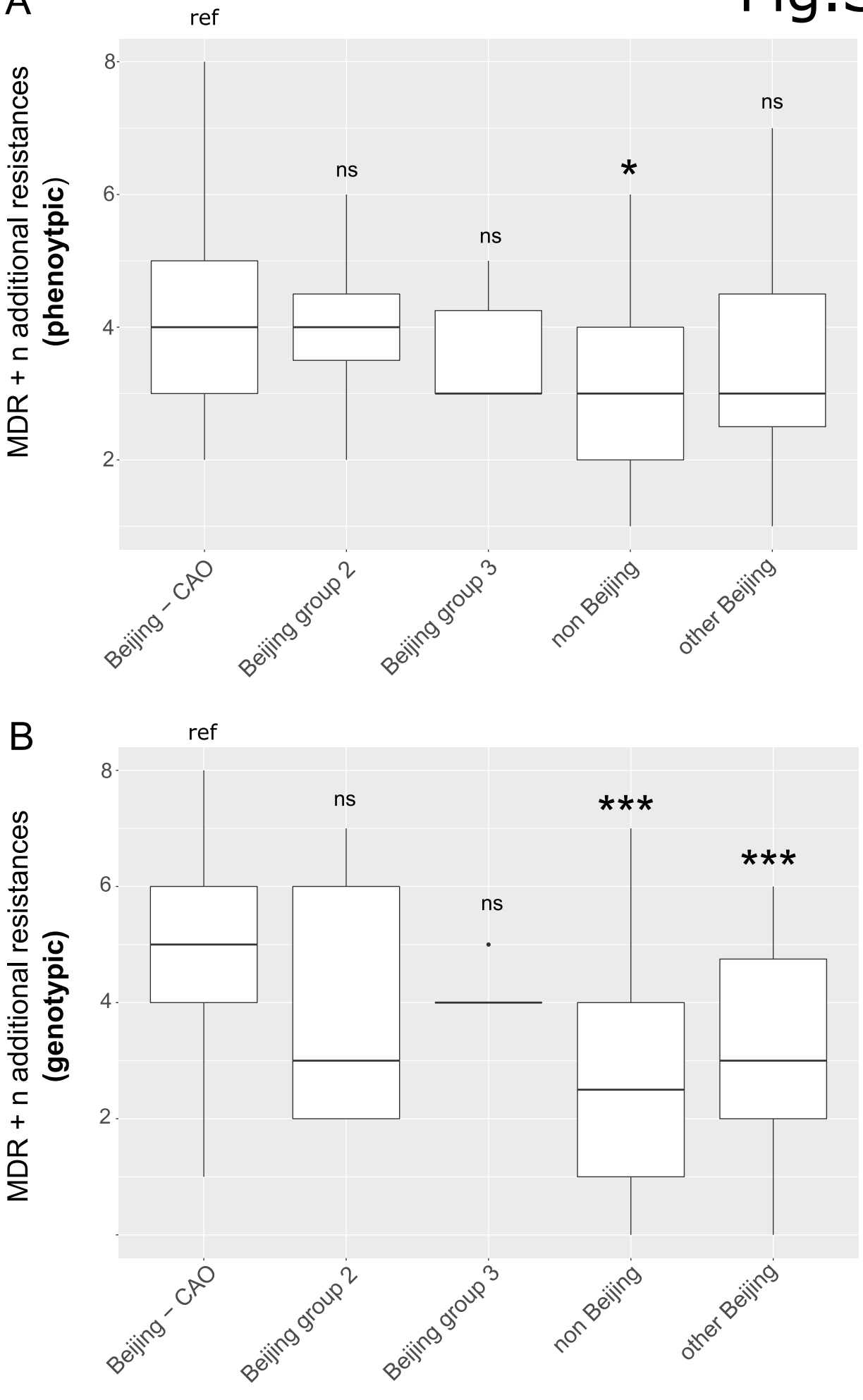

A
ref
ns
ns
*
ns
MDR + n additional resistances
(phenoytpic)
8
6
4
2
Beijing – CAO
Beijing group 2
Beijing group 3
non Beijing
other Beijing
B
ref
ns
ns
***
***
MDR + n additional resistances
(genotypic)
8
6
4
2
Beijing – CAO
Beijing group 2
Beijing group 3
non Beijing
other Beijing

Fig.S4

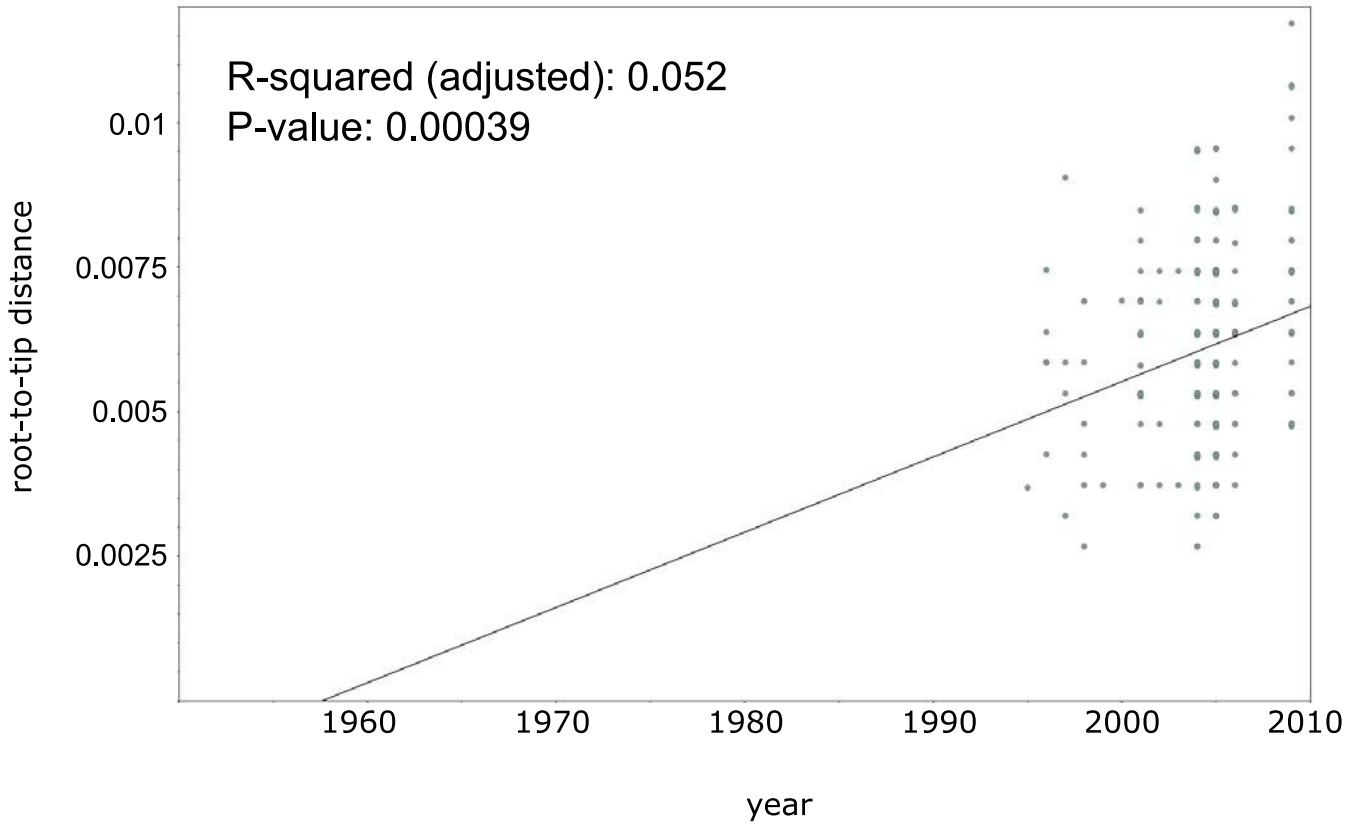

R-squared (adjusted): 0.052
P-value: 0.00039
root-to-tip distance
0.01
0.0075
0.005
0.0025
1960
1970
1980
1990
2000
2010
year

Fig.S5

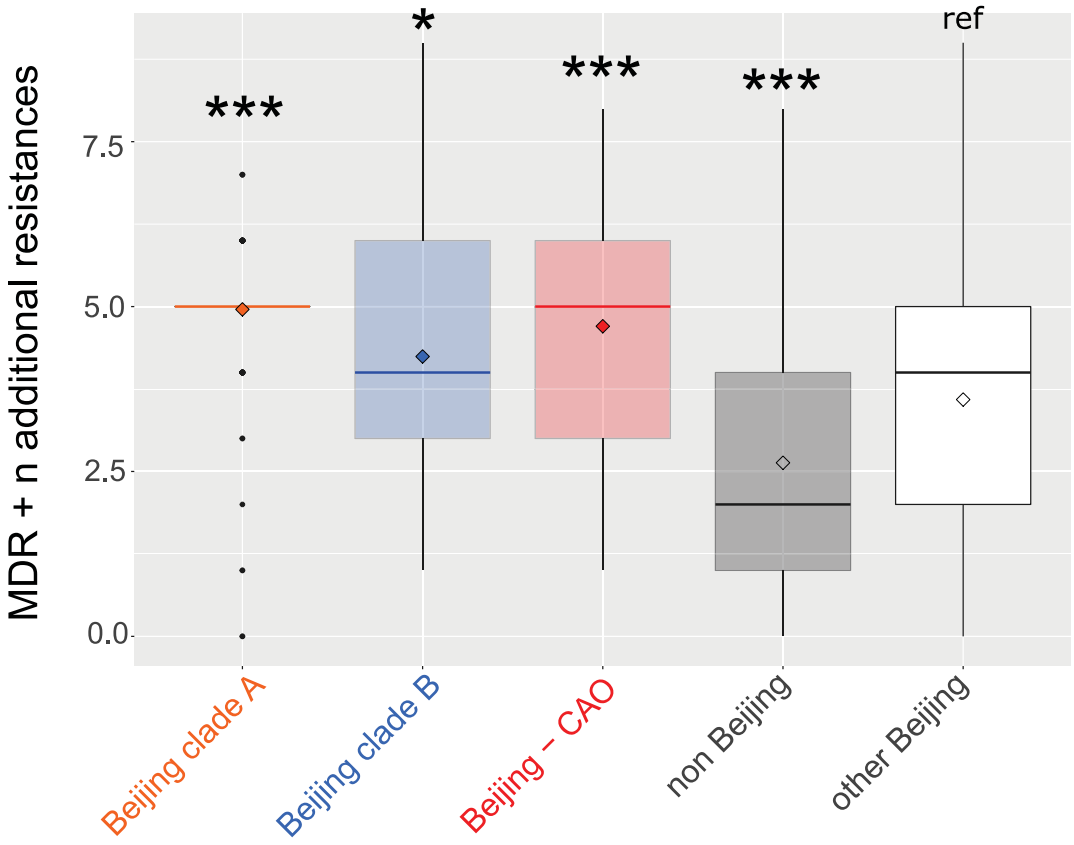

MDR + n additional resistances
7.5
5.0
2.5
0.0
***
*
***
***
ref
Beijing clade A
Beijing clade B
Beijing – CAO
non Beijing
other Beijing

Fig.S6

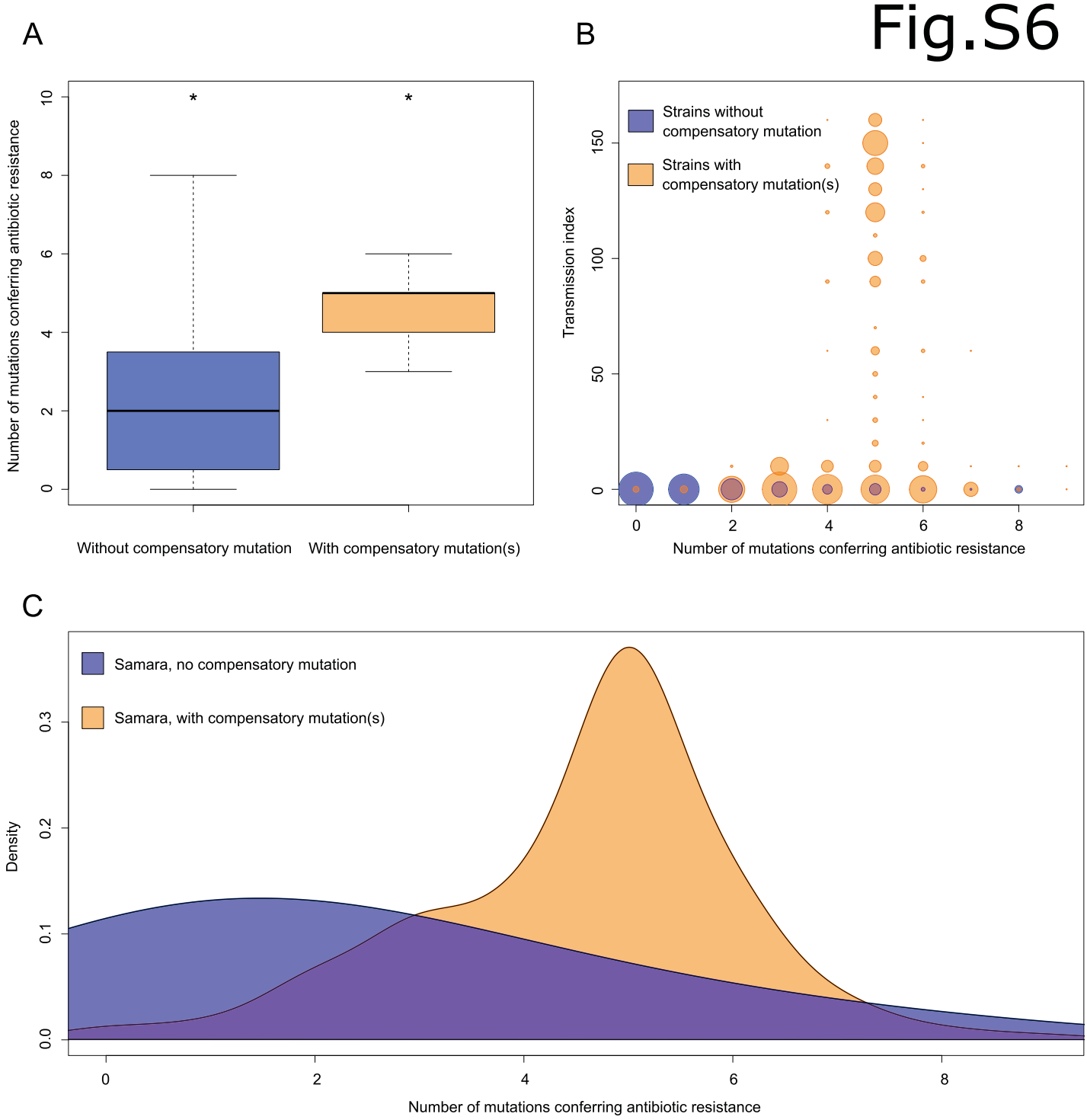
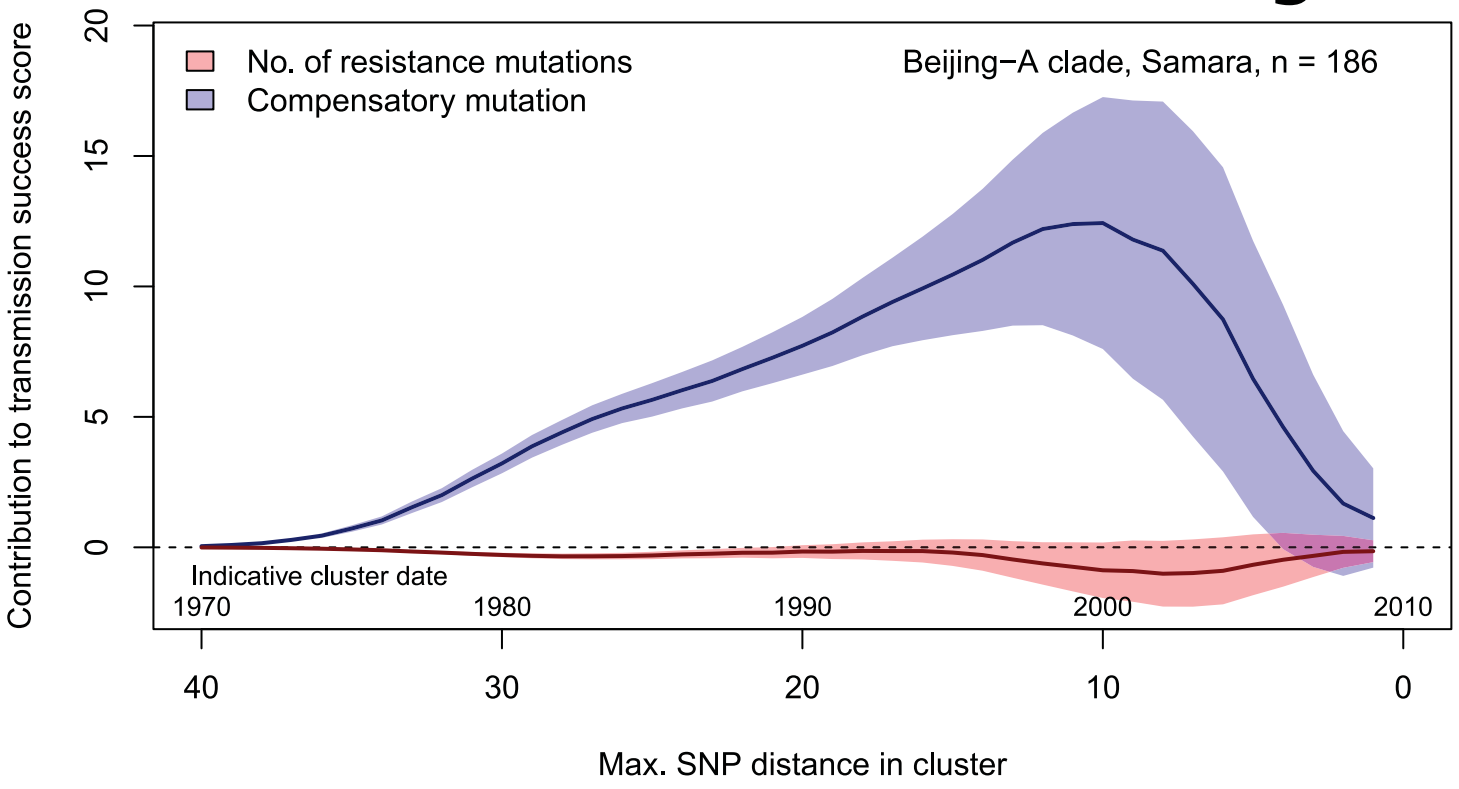

A
B
C
Number of mutations conferring antibiotic resistance
Without compensatory mutation
With compensatory mutation(s)
Transmission index
Strains without compensatory mutation
Strains with compensatory mutation(s)
Number of mutations conferring antibiotic resistance
Samara, no compensatory mutation
Samara, with compensatory mutation(s)
Density
Number of mutations conferring antibiotic resistance

Fig.S7


No. of resistance mutations
Compensatory mutation
Beijing−A clade, Samara, n = 186
Contribution to transmission success score
0
5
10
15
20
Indicative cluster date
1970
1980
1990
2000
2010
40
30
20
10
0
Max. SNP distance in cluster

Chapitre 3.

Investigations sur le succès et le développement de W148, un clone hautement résistant de la famille Beijing de *Mycobacterium tuberculosis*, utilisant des données génomiques

RESEARCH ARTICLE

Evolutionary history and adaptive landscape of the multidrug resistant W148 Russian clone

Authors: Maxime Barbier[3,4*], Matthias Merker[1,2*], [Dr.] Helen Cox[5], Jean-Philippe Rasigade[3,4,6], Silke Feuerriegel[1,2], Thomas A. Kohl[1,2], Egor Shitikov[7], Kadri Toit[8], Roland Diel[9], Sonia Borrell[10,11], Sebastien Gagneux[10,11], Vladyslav Nikolayevskyy[12,13], Sönke Andres[14], Valeriu Crudu[15], Ulrich Nübel[16,17], Philip Supply[18,19,20,21], Stefan Niemann[1,2#] Thierry Wirth[3,4#]

Affiliations:

[1]Molecular and Experimental Mycobacteriology, Research Center Borstel, Germany.

[2]German Center for Infection Research, Borstel site, Germany.

[3]Laboratoire Biologie Intégrative des Populations, Evolution Moléculaire, Ecole Pratique des Hautes Etudes, PSL University, Paris, France.

[4]Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Université Pierre et Marie Curie, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France.

[5]Division of Medical Microbiology and Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa.

[6]CIRI INSERM U1111, University of Lyon, Lyon, France.

[7] Federal Research and Clinical Centre of Physical-Chemical Medicine, Moscow, Russian Federation

[8] SA TUH United Laboratories, Mycobacteriology, Tartu, Estonia

[9] Institute for Epidemiology, Schleswig-Holstein University Hospital, Kiel, Germany

[10] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland.

[11] University of Basel, Basel, Switzerland.

[12] Imperial College London, United Kingdom.

[13] Public Health England, London, United Kingdom

[14] Division of Mycobacteriology (National Tuberculosis Reference Laboratory), Research Center Borstel, Borstel. Germany

[15] National TB Reference Laboratory, Phthysiopneumology Institute, Chisinau, Republic of Moldova

[16]Microbial Genome Research, Leibniz-Institut DSMZ- Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig, Germany.

[17] German Center for Infection Research, Braunschweig site, Germany.

[18] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 8204 - CIIL - Centre d'Infection et d'Immunité de Lille, F-59000 Lille, France.

[19] Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 8204, Center for Infection and Immunity of Lille, Lille, France.

[20] Université Lille Nord de France, Center for Infection and Immunity of Lille, Lille, France.

[21] Institut Pasteur de Lille, Center for Infection and Immunity of Lille, Lille, France.

* contributed equally

# These authors contributed equally to this work. Correspondence to: Prof. Dr. Stefan Niemann (sniemann@fz-borstel.de) and Prof. Dr. Thierry Wirth (wirth@mnhn.fr)

Abstract :

Mycobacterium tuberculosis is a major human pathogen and still belongs, according to the WHO, to the top 10 causes of death in 2015, being the sole microorganism on the list. Yet, all MTBC lineages are not equal, some of them, especially the Beijing one, harbour strains with higher resistance profiles that go through unprecedented waves of propagation. Here, we leveraged a unique, genetically diverse strain collection of the clade B0/W148, also called “Russian clone”, a twig emerging from the Beijing clade, that significantly contributes to MDR epidemics in Russia and Eastern Europe, to determine its evolutionary history and reasons behind its success. Genome sequencing and phylogenetic analyses of 731 isolates, covering the full Eurasian distribution of the clone, highlighted a deep split that confirmed a central Asian origin some 250 years ago. Bayesian demogenetic analyses confirmed the success of W148, revealing two successive expansions, a first one in the late seventies, followed by a second one in the late eighties. Interestingly, the demographic surge and the geographic westward spread were accompanied by numerous adaptive mutations that sharply differentiate W148 from its direct progenitor. This translates into compensatory mutations acquisition without detectable fitness costs and higher transmissibility, sharp antigenic set-up change, amino-acid replacements in type VII secretion system genes and modulation in DNA damage response. These swift changes are the scars of the selective pressures imposed by the host and contribute to our understanding of the adaptive landscape of a successful multidrug-resistant lineage.

Introduction

Mycobacterium tuberculosis the causative agent of tuberculosis, caused 10.4 million new infections in 2015, killing 1.4 million people (WHO 2016). Even though the death toll is decreasing over the years, tuberculosis still remains one of the top ten causes of death worldwide. Tuberculosis is particularly widespread in countries with high HIV prevalence such as South Africa and Nigeria, as well as in countries of high population densities (India, China, Pakistan and Indonesia). Among the different threats to global TB control, drug resistant *Mycobacterium tuberculosis* strains pose a worrying challenge. Currently, five percent of the cases are related to MDR strains, which are resistant to two first line agents, isoniazid and rifampicin, therefore leading to serious societal and medical issues. Among the high burden countries, there are not only the countries mentioned above, but also the Russian Federation and numerous former members of the Soviet Union in Central Asia and Eastern Europe. In Eurasia, MDR strains are, to a large extent, related to the Beijing lineage, which is suspected to acquire mutations conferring drug resistance faster than other lineages (Ford et al. 2013b). In addition to drug resistance conferring mutations some strains may carry compensatory mutations, enabling them to bring several resistances with reduced fitness cost (Handel et al. 2006). Some usual suspects are well known, such as rifampicin resistance compensatory mutations in *rpoA, rpoB* and *rpoC* (Brandis et al. 2012). Indeed, in vitro experiments convincingly demonstrated that strains carrying mutations in one of these genes displayed an enhanced fitness compared to the wild type (Comas et al. 2011). A few studies highlighted the association between these mutations and MDR strains (De Vos et al. 2013; Li et al. 2016), however we still face a lack of proof for a greater transmissibility of strains harboring compensatory mutations in epidemiological studies.

The MTBC is geographically structured (Gagneux and Small 2007; Wirth et al. 2008; Barbier and Wirth 2016) and the same holds for the Beijing lineage where several “clonal-complexes” are characterized. Each of them is associated to a specific geographic area, is off different age and displays a distinct level of success (Merker et al. 2015). A major split segregates “ancient” versus “recent” lineages; the latter being more virulent (Ribeiro et al. 2014). One of the youngest lineages, clade B0/W148, also called “Russian clone”, significantly contributes to the MDR epidemic in Russia and Eastern Europe. This clone, easily identifiable by a large chromosomic deletion (Shitikov et al.

2014) is suspected to have emerged in Siberia, followed by a secondary westward dissemination in the sixties after the breakdown of the gulag system in former Soviet Union (Mokrousov 2013). Since the USSR's fall, W148 strains have become more threatening than ever, causing medical migrations from eastern Europeans to Western Europe seeking treatment against multi drug resistant strains (Faustini et al. 2006). Moreover, it has been proven that W148 plays a prominent role in spinal tuberculosis, a very damaging form of extra-pulmonary TB in Russian patients (Vyazovaya et al. 2015). Nevertheless, until now, few epidemiological surveys focused on clade B0/W148; and most of them relied on MIRU typing, while none developed an extensive population genomic approach.

In order to unravel the geographical source, spread and evolutionary history of the W148 clade, we collected a large dataset composed of 720 W148 strains completed with an additional 11 strains representing their closest relatives found, though lacking the large chromosomic deletion. Isolates were collected between 1995 and 2013, covering a large geographic area, from Western Europe to Central Asia, representing the natural range of the Russian clone. We then focused on the mechanisms involved in drug resistance acquisition in those MDR strains, as well as on the role and impact of known compensatory mutations in the strains global fitness. In a last step, other lineage specific mutations that might explain the success and virulence of W148 were explored.

Results

W148 phylogeography and spatio-temporal dispersion

The dataset comprises 731 strains collected from 24 countries, which are grouped into 15 geographical regions based on a statistical rationale. The vast majority of the collection was sampled during the 21st century (Fig. S1) and essentially covered Western-Europe, Estonia, Belarus and Russia (Fig. 1). After preliminary filtering, sequences were aligned against the H37rv reference genome. This step was followed by the removal of 610 variants located in 67 genes and upstream regions associated with drug resistance and bacterial fitness. A total of 5,264 high-confidence single nucleotide polymorphisms (SNPs) were detected; this number dropped to 4,217 polymorphic sites when the reference genome was excluded and to 3,508 unique SNPs when considering the sole W148 clade. Likelihood mapping analyses (Nieselt-Struwe and von Haeseler 2001) indicated a robust phylogenetic

signal (>78%), albeit with minor occurrence of star-likeness signaling that the tree is well resolved in certain parts only.

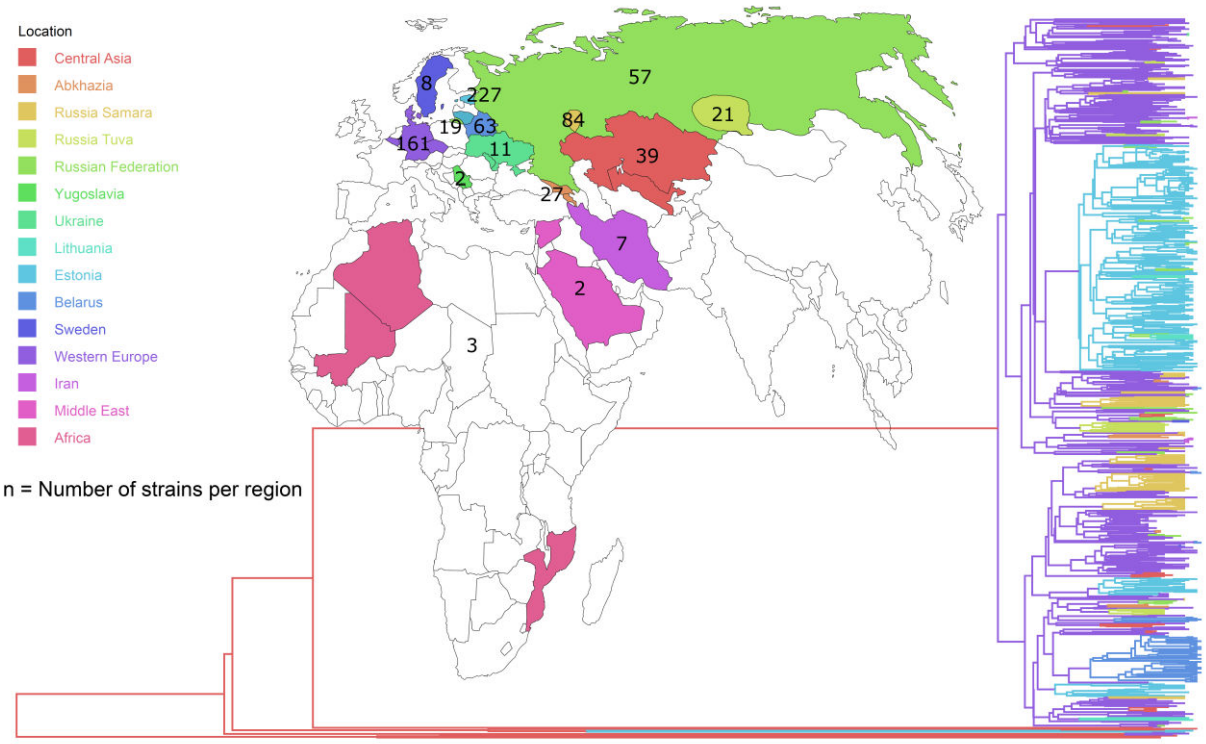


Fig. 1: Bayesian time tree of the 731 W148 strains. All locations were grouped in 15 regions, represented by a panel of different colors following a latitudinal gradient. Branch colors indicate the highest probability of source location. Phylogeny and ancestral locations were inferred using BEAST 2.3.2 (Bayesian skyline model using tip dating and a HKY substitution model, leading to a substitution rate of 1.12 x 10^{-7} substitutions per nucleotide per year) and the GEO SPHERE package. The final tree was generated with FIGTREE 1.4.2.

Before implementing Bayesian phylogenetic reconstructions, a critical step consists in checking if the clone of interest is a measurably evolving population. For this purpose, two independent methods were implemented. First, a plot of genetic distances from a common ancestor against strains sampling time was generated (r^2 = 0.2038, P < 0.001 with 5,000 replicates), providing a strong support for measurable accumulation of DNA sequence variation over the sampling time interval (Fig. S2). Second, marginal likelihood scores of two alternative Bayesian models implementing respectively a fixed mutation rate and a tip dating approach were compared. According to BEAST, the tip-dating model was clearly favored ($\log_{10}$ Bayes factor = 14.831). Furthermore, in order to accurately estimate the nucleotide substitution rate, different clock and demographic models were compared (Table 1). The best fitting evolutionary model was obtained under a Bayesian skyline model with a relaxed clock, leading to 1.15 x 10^{-7} mutation per position per year, with 57.61% of variation across the tree.

Table 1: Comparisons of different demographic and clock models under BEAST

Strains	Clock model	Demographic model	Substitution rate (SNPs/nucleotide/year)	TMRCA	Likelihood	BF (clock model)	BF (demographic model)
731	Fix	Constant	9.94983E-08	245 (222-266)	-44182		
731	Tip dating, Strict clock	Constant	1.02425E-07	247 (214-281)	-44151	62	
731	Tip dating, Relaxed clock	Constant	1.14599E-07	285 (183-402)	-43947	408	
720	Tip dating, strict clock	Constant	1.02542E-07	55 (47-65)	-39430		
720	Tip dating, relaxed clock	Constant	1.16706E-07	68 (48-97)	-39215	430	
720	Tip dating, strict clock	Skyline	9.10702E-08	49 (39-59)	-39389		82
720	Tip dating, relaxed clock	Skyline	1.11906E-07	39 (29-51)	-39191	396	48

The Bayesian time tree underscores a clear distinction between the 11 basal central Asian strains and the W148 outbreak strains, the latter being divided into two major sublineages (Fig. 1). It is worth mentioning that Western European strains were equally distributed in the phylogeny, whereas the Estonian strains first and foremost clustered in a single clade. The epidemic character of the Estonian strains is well depicted in a minimum spanning tree (Fig. S3) where central nodes containing several identical strains are surrounded by numerous 2-5 SNPs variants. The same holds but to a lesser extend for the Belarus strains. In sharp contrast, European isolates are randomly distributed in the network. Next, to reconstruct the source and dissemination routes of the strains, we performed a phylogeographic analysis using an ancestral state reconstruction model in BEAST v2.3.2. According to the Bayesian model, the most likely origin of the root and basal branches sticks in Central Asia (Fig. S4 and Fig. S5). The likely origin of the W148 clade is Western Europe, with a most recent common ancestor inferred in the early sixties. The time tree indicates a sharp dispersion from the early seventies in all Eurasia, followed by two main outbreaks in Belarus and Estonia in the late eighties. According to the coalescent-based demographic reconstructions, the clone went through two expansions, a first twentyfold increase in the late seventies, followed by a secondary fivefold increase in the late eighties (Fig. S6).

W148 Genetic diversity

In order to correct for unequal sample sizes and to generate an unbiased strain genetic diversity picture we implemented a rarefaction procedure and computed the mean π's for each region as described by Nei and Li (Nei and Li 1979). Bootstrapping and subsampling procedures where adjusted to the number of strains corresponding to the region with the lowest coverage (Lithuania, N = 19 strains). Regions displaying the highest nucleotide diversity are located in Central Asia, Western Europe and Russia (Fig. S7) with respectively 7.19 x 10^{-3}, 7.14 x 10^{-3} and 6.33 x 10^{-3} mean π values. Nevertheless, Western Europe is characterized by a great variance, whereas Tuva's π estimate in Southern Siberia only reached 4.08 x 10^{-3}. However, we should keep in mind that the Western-European sample is not necessary representative of the local population since patient's origin was unknown in 63% of the cases. For those whose origin was documented, only 34% came from Western Europe, whereas the bulk of them came from Russia (37%) and other Eastern European and Central Asian countries.

Strains resistance profile

Proportions of genotypic resistance for nine antibiotics across the dataset were as follows (Fig. 2): Isoniazid (99.6%), streptomycin (99.5%), rifampicin (92.5%), ethambutol (84.0%), pyrazinamide (49.4%), fluoroquinolones (22.8%), kanamycin (63.2%), thioamide (41.7%), PAS (9.2%). Overall, 80.1% of the isolates carried a compensatory variant, 45% were classified as pre-XDR (either fluoroquinolone or second-line injectable drug resistance related variants) and an additional 20% classified as XDR-TB. We then compared the profile of resistance between the different regions by performing chi-square tests and logistic regressions. Belarus was the region with the highest drug resistance rates, showing significantly more resistant strains ($P < 0.01$) than expected for amikacin, capreomycin, fluoroquinolones, TAs and pyrazinamide, and the largest fraction of XDR strains (Fig. S8). In contradistinction, Western Europe and Estonia harbored less resistant strains than expected ($P < 0.01$) to ethambutol, fluoroquinolones, kanamycin, amikacin, capreomycin, and TAs.

Impact of compensatory mutations

We then focused on the impact of potential compensatory mutations on strain fitness. Considering three genes, respectively *rpoA, rpoB* and *rpoC*, over 731 isolates, 141 did not harbor any compensatory mutations, 580 harbored mutations in one of them, whereas only 10 exhibited

compensatory mutations in at least two distinct genes. As shown in Figure 2, compensatory mutations were mostly confined to *rpoC*.

Next we compared the number of resistance mutations and, as a proxy of fitness, the transmission index between strains carrying compensatory mutations against those who did not. Interestingly, the mean number of resistance mutations was higher among strains carrying compensatory mutations (Fig. 3A), 5.25 vs 3.55 (Welch t-test, $P < 10^{-14}$).

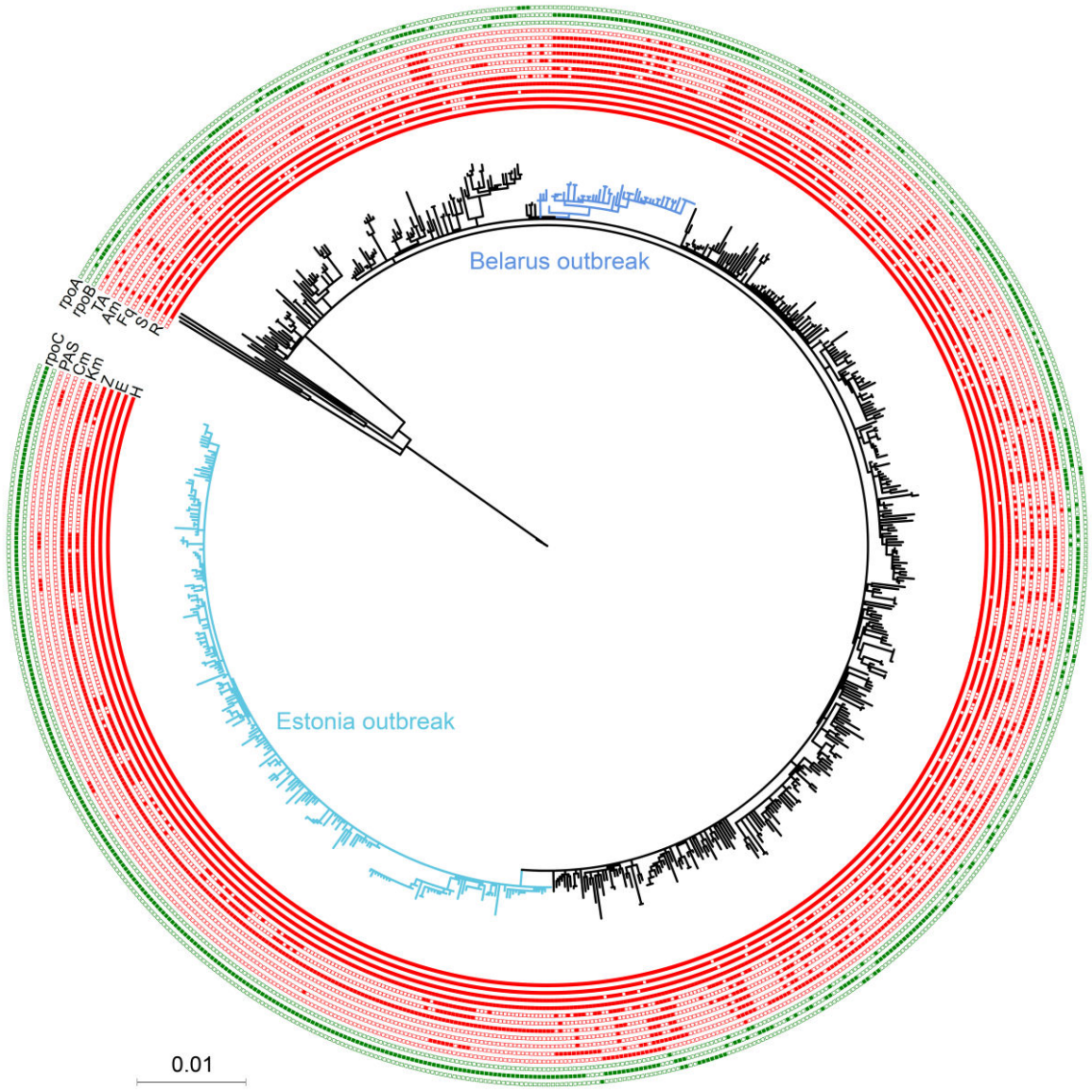


Fig. 2: Maximum Likelihood tree based on 5,264 variable positions (SNPs) among 731 isolates originated from 15 regions obtained using PHYML (HKY85 substitution model) and visualized with ITOLV.4 (LETUNIC AND BORK 2016). The Belarus and Estonian outbreaks are highlighted. The different crowns are showing the presence (filled box) or absence (empty box) of mutations conferring resistance to different drugs: H=Isoniazid, R=Rifampicin, E=Ethambutol, S=Streptomycin, Z=Pyrazinamide, Fq=Fluoroquinolones, Km=Kanamycin, Am=Amikacin, Cm=Capreomycin, TA=Thioamide and PAS=Para-aminoslicylic acid (in red); and presence or absence of compensatory mutations: rpoA, rpoB and rpoC (in green).

Moreover, in the same vein, strains with compensatory mutations also showed larger transmission indices than strains presenting no compensatory mutation, 17.39 vs 10.05 ($P = 4 \times 10^{-5}$) (Fig. 3B). To disentangle the relative contributions of resistance and compensatory mutations to the transmission success, those characteristics were included in multiple regression models of transmission indices with varying SNP cutoffs, equivalent to moving the time back to the MRCA (i.e., the inferred date of emergence) in a sliding window of transmission networks (Fig. 3C). While controlling for compensatory mutations, number of resistance mutations was negatively associated with transmission network size, especially at a 30 years scale, suggesting that the fitness cost of resistance was greater than the benefits obtained from multiple resistance profiles in W148 strains. On the other hand, compensatory mutations were positively associated with transmission network size, independently of the number of resistance mutations, peaking at a 15 years scale, indicating a higher transmissibility for isolates harboring such compensatory mutations.

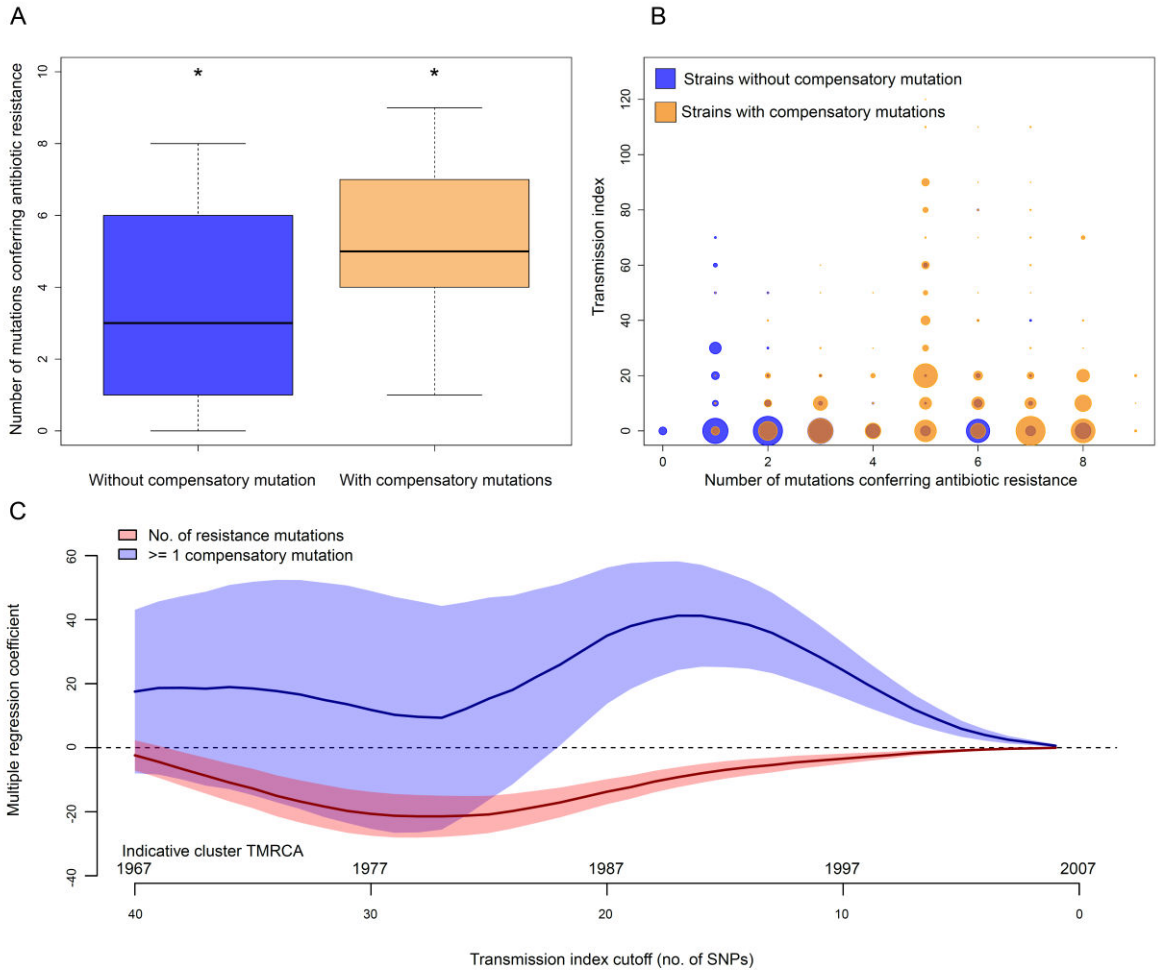


Fig. 3: Comparisons between strains carrying compensatory mutations (in orange) and strains without compensatory mutations (in blue). A) Boxplot showing number of resistance mutations for the two categories (without or with compensatory mutations). The two categories were significantly different (Welch's t-test, $P < 2.2 \times 10^{-16}$. B) Bubble plots showing the transmission index (number of strains differing by less than 10 SNPs) as a function of antibiotic resistance related mutations. Bubble sizes are function of the number of strains. C) Relative contributions of resistance associated and compensatory mutations to the transmission success of W148. Transmission indices with varying SNP distance cutoffs (abscissa) were computed to represent different transmission periods. Based on MTBC genome evolution rate, a cluster of strains diverging by at most x SNPs has TMRCA ≈ x years (dates indicated above the abscissa). For each cutoff value, multiple linear regression model of the transmission index was constructed. Shown are the coefficients (solid lines) and 95% confidence interval (colored bands) of the no. of resistance mutations (red) and the presence of putative compensatory mutations (blue). Coefficients are significant at the 5% level for cutoff values where the confidence band does not include zero (dashed line).

Accumulation of resistance and compensatory mutations

Another important topic concerns the dynamic and the accumulation order of resistance and compensatory mutations both in quantitative and qualitative manners. By performing regression analyses, we were able to detect an accumulation of genotypic resistances (in red) and compensatory mutations (in blue) over time (Fig. 4) with $r^2 = 0.09$; $P < 2.2 \times 10^{-16}$ and $r^2 = 0.11$; $P < 2.2 \times 10^{-16}$ respectively.

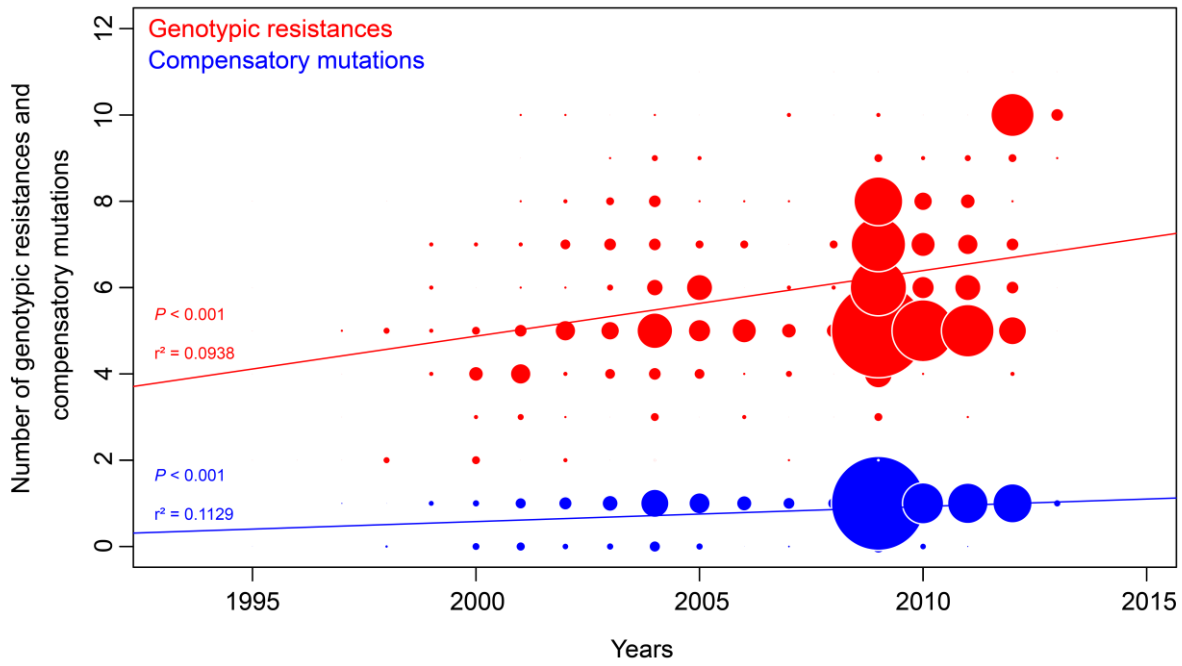


Fig. 4: Bubble plot showing the number of genotypic resistances (in red) and compensatory mutations (in blue) as a function of strain's years of isolation. Bubble sizes are proportionate to the number of strains. Regression analyses were significant ($P < 2.2 \times 10^{-16}$) and regression curves are shown on the plot.

From a qualitative point of view, we scrutinized the paired appearance of mutations conferring resistance and compensatory effects (Fig. 5). As expected, resistance mutations to first line drugs such as rifampicin, isoniazid, streptomycin and ethambutol were the most common mutations. They occurred more frequently in first rank than second, compared to other mutation types. Mutations offering resistances to new TB drugs such as delamanid, PA-824, bedaquiline and clofazimine or drugs used to treat MDR strains such as cycloserine and linezolide were almost exclusively secondary. The most frequent compensatory mutations affected *rpoC* and appeared as often in first or second rank. Disparately, *rpoA* and *rpoB* SNP's rarely seemed to precede other mutations and thus are likely to represent the latest step in the resistance profile evolution of these strains.

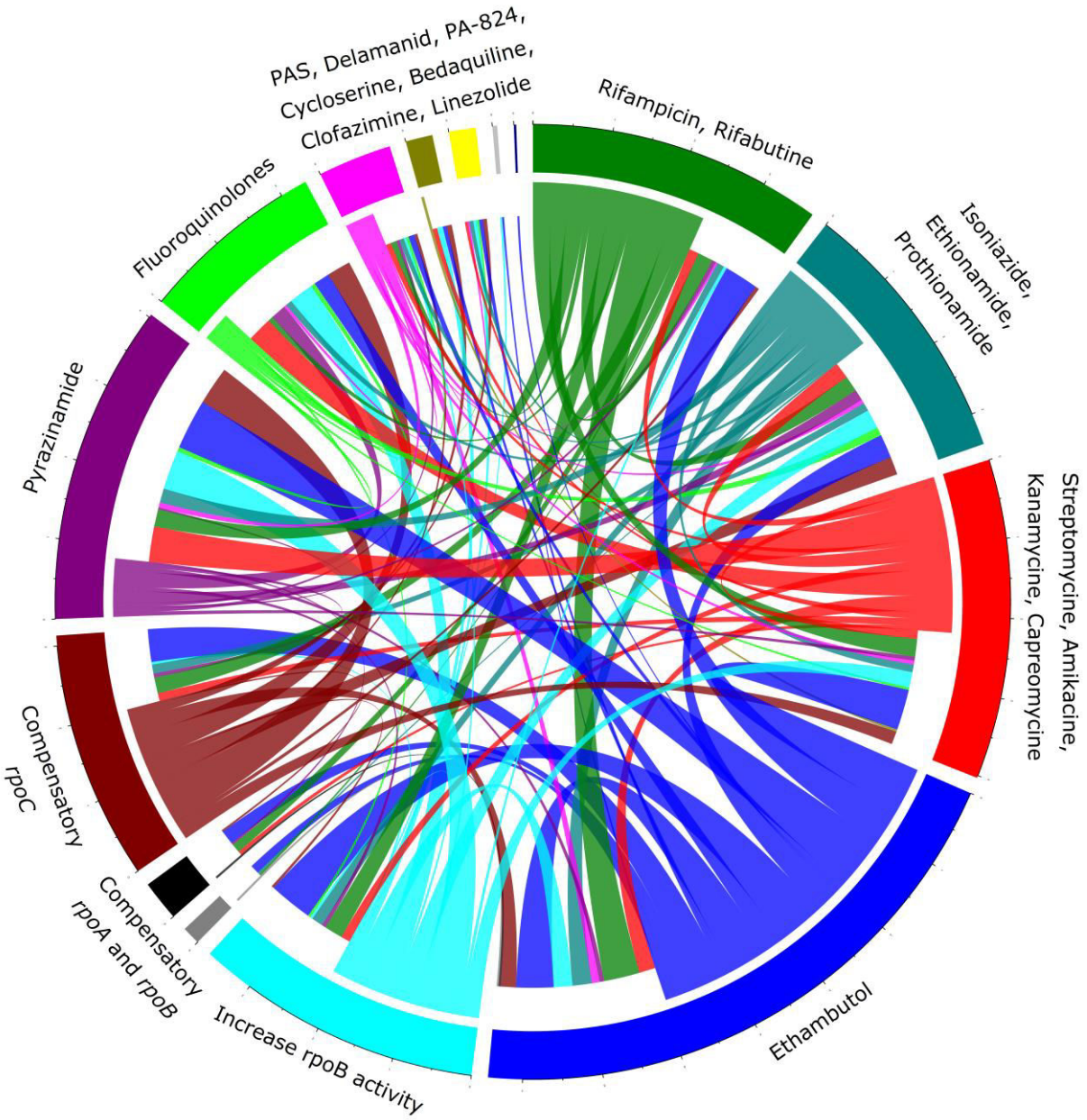


Fig. 5: Representation of the resistance and compensatory mutations occurrences by pair. On the crown, are shown the different mutations, grouped in categories according to the drug resistances they are conferring. More a drug occupies an important part of the circle and more the mutations offering a resistance to this antibiotic are numerous. Link between the crowns indicates that one mutation occurred just before a second one. The direction of the association is indicated by the color of the link and its proximity to the crown. The width of the link indicates the frequency of the association. For example, a green link between "Rifampicin, Rifabutin" and "Ethambutol", indicates that several times, a mutation conferring resistance to the rifampicin or the rifabutin occurred before a mutation conferring ethambutol's resistance. The width of the link indicates the number of times this association occurred comparatively to others. To enhance visibility we removed singleton and doubleton (association that occurred only one or two times).

PCAdapt and W148 adaptive signature

In the frame of the Bayesian principal component analysis, principal component 2 segregated the 720 W148 outbreak strains from the 11 basal strains (Fig. S9). SNPs associated with PC2 and displaying the highest Mahalanobis distance are shown in Table 2.

Table 2: List of the W148 specific SNP's under potential positive selection according to PCADAPT

Original SNP position	-log10(p-value) of the Mahalanobis distance	Max correlated PC	Position in the *H37rv* genome	Ref	Gene	Annotation
24	218273.1149	2	4086	G	*Rv0003*	DNA replication and repair protein RecF
229	218273.1149	2	143120	C	*Rv0118c*	Probable oxalyl-CoA decarboxylase OxcA
539	218273.1149	2	404130	T	*Rv0338c*	Probable iron-sulfur-binding reductase
1280	218273.1149	2	1023883	G	-	-
1291	218273.1149	2	1035426	T	*Rv0928*	Periplasmic phosphate-binding lipoprotein PstS3
1471	218273.1149	2	1155884	T	*Rv1031*	Probable potassium-transporting ATPase C chain KdpC
1524	218273.1149	2	1203693	C	*Rv1078*	Probable proline-rich antigen homolog Pra
1607	218273.1149	2	1272899	T	*Rv1145*	Probable conserved transmembrane transport protein MmpL13a
1845	218273.1149	2	965248	A	*Rv0867c*	Possible resuscitation-promoting factor RpfA
1961	218273.1149	2	1268475	G	*Rv1141c*	Probable enoyl-CoA hydratase EchA11
2554	218273.1149	2	564588	C	*Rv0473*	Possible conserved transmembrane protein
2708	218273.1149	2	2487024	C	*Rv2219A*	Probable conserved membrane protein
2796	218273.1149	2	1094734	A	*Rv0979c*	Hypothetical protein
3102	218273.1149	2	2517129	A	*Rv2243*	Malonyl CoA-acyl carrier protein transacylase FabD
3139	218273.1149	2	2544135	A	*Rv2268c*	Probable cytochrome P450 128 Cyp128
3165	218273.1149	2	2569288	G	*Rv2298*	hypothetical protein
3568	218273.1149	2	2908357	G	*Rv2583c*	Probable GTP pyrophosphokinase RelA
3676	218273.1149	2	3014343	C	*Rv2698*	Probable conserved alanine rich transmembrane protein
3699	218273.1149	2	3031090	G	*Rv2719c*	Possible conserved membrane protein
4430	218273.1149	2	3697016	G	*Rv3309c*	Probable uracil phosphoribosyltransferase Upp
5009	218273.1149	2	3497859	G	*Rv3132c*	Two component sensor histidine kinase DevS
5167	218273.1149	2	4338371	T	*Rv3862c*	Possible transcriptional regulatory protein WhiB-like WhiB6
5181	218273.1149	2	4349982	G	*Rv3871*	ESX conserved component EccCb1 ESX-1 type VII secretion system protein
5206	218273.1149	2	4367633	T	*Rv3885c*	ESX conserved component EccE2 ESX-2 type VII secretion system protein Possible membrane protein
1027	87006.51827	2	826756	C	*Rv0734*	Methionine aminopeptidase MapA
1107	87006.51827	2	886661	C	-	-
1726	87006.51827	2	1353462	G	*Rv1209*	hypothetical protein
2136	**87006.51827**	**2**	**1402823**	**A**	***Rv1255c***	**Probable transcriptional regulatory protein**
2685	**87006.51827**	**2**	**2241091**	**C**	***Rv1997***	**Probable metal cation transporter P-type ATPase A CtpF**
2900	87006.51827	2	3070325	C	*Rv2757c*	Possible toxin VapC21
3767	87006.51827	2	3086731	A	*Rv2779c*	Possible transcriptional regulatory protein (probably Lrp/AsnC-family)
4762	87006.51827	2	1516145	G	*Rv1349*	Iron-regulated transporter IrtB
5022	87006.51827	2	3799523	G	*Rv3384c*	Possible toxin VapC46 Contains PIN domain
5023	87006.51827	2	3798062	G	*Rv3383c*	Possible polyprenyl synthetase IdsB
5024	87006.51827	2	4221565	G	*Rv3776*	hypothetical protein
5025	87006.51827	2	4221586	A	*Rv3776*	hypothetical protein
5026	87006.51827	2	4221591	C	*Rv3776*	hypothetical protein
5027	87006.51827	2	4221609	G	*Rv3776*	hypothetical protein
5028	87006.51827	2	4221619	G	*Rv3776*	hypothetical protein
5152	87006.51827	2	4321479	G	-	-
950	58406.48838	2	751999	C	*Rv0655*	Possible ribonucleotide-transport ATP-binding protein ABC transporter MkI
1119	58406.48838	2	895164	G	*Rv0802c*	Possible succinyltransferase in the GCN5-related N-acetyltransferase family
1613	58406.48838	2	1276392	G	*Rv1148c*	hypothetical protein
2057	58406.48838	2	1377129	T	*Rv1234*	Probable transmembrane protein
2277	58406.48838	2	4371444	C	*Rv3887c*	ESX conserved component EccD2 ESX-2 type VII secretion system protein Probable transmembrane protein
2410	58406.48838	2	3370081	A	*Rv3011c*	Probable glutamyl-tRNA(GLN) amidotransferase (subunit A) GatA (Glu-ADT subunit A)
2615	58406.48838	2	784778	A	-	-
3627	58406.48838	2	2964454	C	*Rv2638*	hypothetical protein
3910	58406.48838	2	3203157	C	*Rv2893*	Possible oxidoreductase
4278	58406.48838	2	3557244	C	-	-
4287	58406.48838	2	3566107	C	*Rv3196*	hypothetical protein
4338	58406.48838	2	3610335	G	*Rv3233c*	Possible triacylglycerol synthase (diacylglycerol acyltransferase)
1071	58286.9157	2	854604	C	*Rv0760c*	hypothetical protein
1441	58286.9157	2	1130195	C	*Rv1011*	Probable 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase IspE (CMK)
2156	58286.9157	2	4167689	T	*Rv3722c*	hypothetical protein
2977	58286.9157	2	2392289	G	*Rv2130c*	Cysteine:1D-myo-inosityl 2-amino-2-deoxy--D-glucopyranoside ligase MshC
3009	58286.9157	2	2422061	A	*Rv2160A*	hypothetical protein
5087	58286.9157	2	4280254	G	*Rv3815c*	Possible acyltransferase
672	48712.16065	2	527895	T	*Rv0439c*	Probable dehydrogenase/reductase
3872	48712.16065	2	3161858	T	*Rv2852c*	Probable malate:quinone oxidoreductase Mqo (malate dehydrogenase [acceptor])
4951	48712.16065	2	2573793	A	-	-
1137	39785.48207	2	907046	G	*Rv0812*	Probable amino acid aminotransferase
1467	39785.48207	2	1154618	C	*Rv1030*	Probable potassium-transporting P-type ATPase B chain KdpB
2606	39785.48207	2	735672	G	*Rv0641*	50S ribosomal protein L1 RplA
3205	39785.48207	2	2604740	A	*Rv2331A*	Hypothetical protein
3617	**39785.48207**	**2**	**2954197**	**G**	***Rv2627c***	**hypothetical protein**
3797	39785.48207	2	3109512	G	*Rv2800*	Possible hydrolase
3563	39746.83614	2	2904110	C	*Rv2579*	Possible haloalkane dehalogenase DhaA (1-chlorohexane halidohydrolase)
5254	37694.39815	2	4399683	C	*Rv3910*	Probable conserved transmembrane protein
2684	26805.96113	2	2691800	G	*Rv2395*	Probable conserved integral membrane protein
1790	23712.28728	2	964177	C	*Rv0866*	Probable molybdenum cofactor biosynthesis protein E2 MoaE2
1305	2.715199159	2	1040491	G	*Rv0932c*	Periplasmic phosphate-binding lipoprotein PstS2 (PBP-2) (PstS2)
1267	0.758636137	2	1008964	A	*Rv0906*	hypothetical protein
1531	0.758636137	2	1209373	C	*Rv1084*	hypothetical protein

Of note is that many non-synonymous SNP's affected antigens or genes displaying epitopes, underscoring the tremendous role of host-pathogen interaction in the short-term evolutionary arm race. Three targeted genes belong to top scores immunogenic antigens (Rv3132c, Rv1031 and Rv1349). The first one, DosS, is a membrane bound sensor histidine kinase involved in dormancy (Gerasimova et al. 2011) that seems to play a major role in the adaptation to hypoxia (Sherman et al. 2001). The second is member of the top 4 candidates out of 34 proteins in a systematic screen seeking for novel cellular and serological antigen biomarkers of *M. tuberculosis* (Liu et al. 2014) and presents a MET11 to THR11 replacement at cell outer surface (Fig. 6). The last one, irtB (top ranking 45 antigens) part of the iron acquisition apparatus is a probable drug-transport ATP-binding protein ABC transporter (Zvi et al. 2008). In the same vein, non-synonymous mutations affected Rv3862c an immunogenic antigen that induces long-term IFN- γ response in the host (Kassa et al. 2012; Serra-Vidal et al. 2014) but also Rv3871, which encodes an FtsK/SpoIIIE protein with two human T cell epitopes. Rv3871 is part of region of difference 1 (RD1) and is required for secretion of both ESAT-6 and CFP-10 (Gao et al. 2004; Brodin et al. 2006; McLaughlin et al. 2007). Though not being located in the T cell epitope area, the ALA386 to SER386 mutation is discovered in a large pocket region found to be the location of active sites (Fig. 6).

Another family of genes impacted by non-synonymous mutations belongs to the nutrition, starvation gene category. Rv0928, Rv3132c and Rv2638 presented amino acid changes; those genes are involved in adaptation to inorganic phosphate depletion (Ferraris et al. 2014), latency and hypoxic conditions (Sherman et al. 2001; Gerasimova et al. 2011), and stress signaling respectiviely. The amino-acid change affecting Rv0928 according to UCSF CHIMERA seems to be in the C-terminal cytoplasmic environment (Fig. 6). Furthermore, two toxic components of type II toxin-antitoxin module harbored W148 specific amino-acid changes, affecting respectively VapC21 and VapC46.

In addition, the SOS-repair category demonstrates an interesting pattern. Rv3776, a probable member of the *M. tuberculosis* 13E12 repeat family that landed into the LexA regulatory system to allow itself to become active when DNA damages occur (Davis et al. 2002), had no less than two distinct amino acid changes. Intriguingly, another DNA-damage-inducible gene (Rv2719c), which is divergently

transcribed relative to *lexA* (Dullaghan et al. 2002), was interrupted by a STOP codon in position 150. It has not escaped our notice that increased levels of Rv2719c are directly responsible for cell division inhibition, reduced growth of *M. tuberculosis* in nutrient broth media, human peripheral blood monocytes and murine macrophage cell line (Chauhan et al. 2006). Yet another replacement targeted Rv2579 (dhaA), another LexA binding region identified by ChIP-seq (Smollett et al. 2012), pointing towards a preponderant role of the SOS response modulation in the Russian lineage.

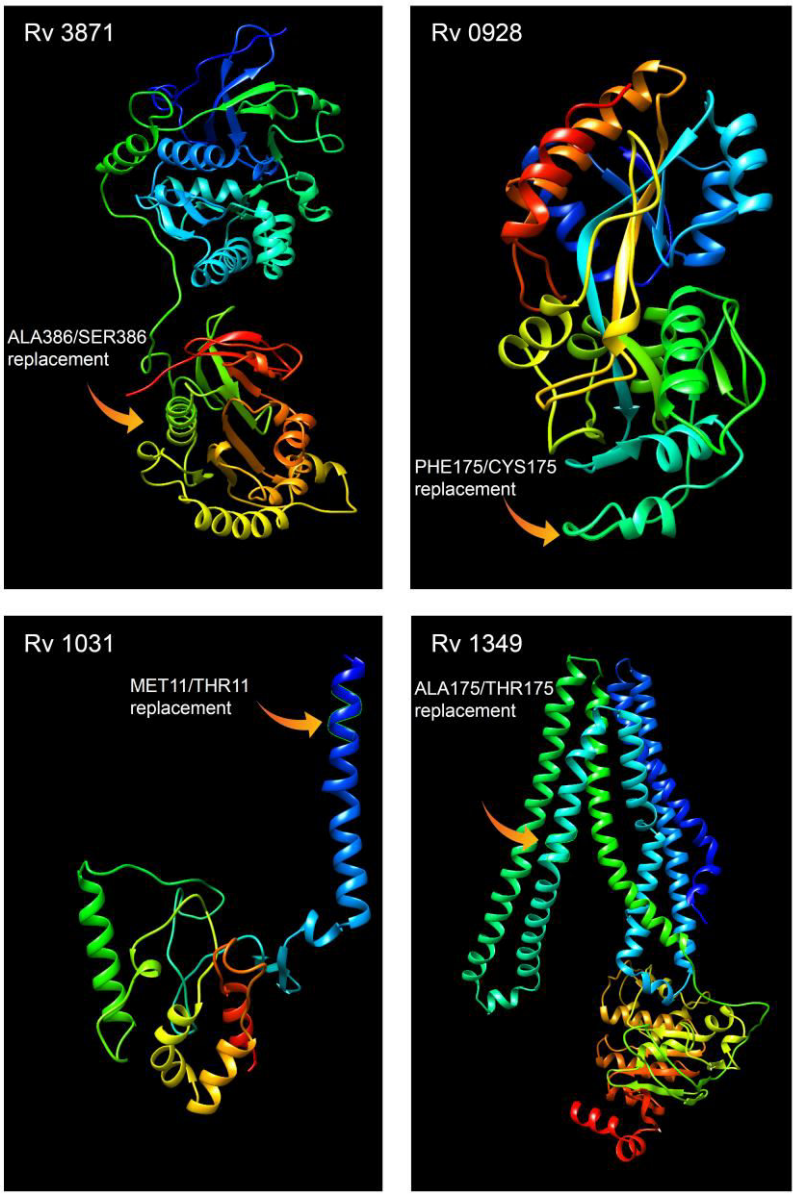


Fig. 6: Positions on the tertiary protein structure of W148 lineage specific non-synonymous mutations for a set of four antigenic genes. Visualization and predictions were executed by PHYRE2 software (http://www.sbg.bio.ic.ac.uk/phyre2).

Homoplastic SNPs

An alternative analysis of the SNPs matrix based on a homoplasy approach spotted two amino-acid changes of potential interest. One of them affected the hddA gene, leading to a unique amino-acid replacement that is possibly involved in a lower enzymatic production compared to non-W148 lineages according to proteomic analyses (Bespyatykh et al. 2016). The second impacted *lppB*, a gene undergoing recombination with more than 40% of the sites being under positive selection (Phelan et al. 2016).

Discussion

We have compiled a large collection of multi-drug resistant *M. tuberculosis* Beijing strains encompassing 720 W148 strains completed with 11 closely related isolates (showing the same MIRU profile) collected through all Eurasia. By performing whole genome sequencing, we were not only able to revisit scenarios gathered from MIRU genotyping, but also unraveled new features concerning the evolutionary history of W148. The substitution rate was estimated at 1×10^{-7} mutation/genome/year for W148, a result which is highly congruent with former empirical studies (Eldholm et al. 2015; Merker et al. 2015; Eldholm et al. 2016) as well as with in vitro estimations (Ford et al. 2011). The calibration of the molecular clock allowed us to confirm the date of appearance of the clone in the early sixties (Mokrousov 2013). As indicated by the BEAST V2.3.2 analyses and regional nucleotide diversity, the geographic region where W148 or at least W148's ancestor emerged was most likely Central Asia, before it disseminated towards the north and the west of Eurasia. At that time, Kazakhstan and Uzbekistan were part of the former USSR. These former soviet republics are well known to be plagued by strains belonging to two Beijing lineages, the Central Asian Clade (Eldholm et al. 2016) and the Clonal Complex 2 (Merker et al. 2015), which includes W148. The low incidence (3.6%) of W148 in this region (Mokrousov 2013) does not necessarily contradict our findings (highest diversity in Central Asia) but could rather be explained by natives' long-term adaptation (Gagneux et al. 2006) and possible competition with other lineages. Furthermore, we cannot exclude the fact that endemic local strains might go through a bottleneck and experience a secondary epidemic phase in a carceral setting. The next steps hypothesized in Mokrousov et al. 2013,

namely that W148 disseminated from West Siberia to North Siberia, Western Russia and Eastern Europe, could not be firmly addressed in this study due to a lack of precise geographic information regarding numerous Russian isolates. Yet, we found a higher genetic diversity in Western Europe than in Russia, and the Bayesian analyses pointed towards Western Europe as the place reached after Central Asia instead of Russia. This paradoxical result is driven by the exceedingly high genetic diversity found in Western Europe, which is likely the product of massive economic and medical migrations. W148 strains are MDR, thus really difficult and expensive to cure, steering infected people to seek treatments in western countries. Meanwhile, some migrants are fleeing war and misery too, which are favorable settings for tuberculosis development. Inexorably they often come from high-incidence areas and have a much higher chance to be infected compared to populations in the West (van der Werf and Zellweger 2016). In the late eighties, W148 dispersed across all Europe, causing two major outbreaks: a first one in Estonia, putting the country among the 27 high multi-drug resistant TB burden countries in the world (WHO 2011), then a second one in Belarus. Those sublineages are exclusively formed by many pre-XDR strains and strains with enhanced resistances for amikacin, capreomycin, fluoroquinolones, TAs and pyrazinamide. This situation is in all probability driven by poor TB management over several decades (Skrahina et al. 2013).

The demographic analyses revealed a strong biphasic expansion of W148. The first massive surge occurred during the mid-seventies just after the publication of the Gulag Archipelago in Europe from Nobel Prize holder Aleksandr Solzhenitsyn. This strong increase in W148 population size does not match major socioecomic events; *i.e.* the massive release of political prisoners in the mid fifties, nor the dissolution of the Gulag system in 1960 by Khrushchev. A plausible causality might be related to the introduction of new antibiotic treatments in the seventies, combining rifampicin and isoniazid, joined by the addition of pyrazinamide a bit later (Iseman 2002). Thanks to its enhanced ability to acquire resistances, the W148 clone could have prospered when other strains, more sensitive to drugs, would have fell. Concomitantly during this period, the population density in Siberian cities reached a maximum, inflating therefore the niche of the bug.

Concerning the second mild growth phase in the early nineties the explanation is more straightforward. Indeed, this secondary increase in effective population size correlates with the fall of

the Soviet Union and its health system (Médecins Sans Frontières, 2011); an episode that led to a 50% increase in the mortality rate. The same onset was observed for other diseases such as diphtheria (Vitek and Wharton 1998) and syphilis (Renton et al. 1998) that reached epidemic peak in Eastern Europe and Russia at the same period.

Yet, what are the intrinsic features that made the success of W148? One of the crucial factors, highlighted in this study, explaining the success of the W148 strains is the acquisition of compensatory mutations. These mutations in *rpoA*, *rpoB* and *rpoC* are known to dramatically decrease the fitness cost associated with rifampicin's resistance (Brandis et al. 2012; Li et al. 2016). As exemplified here, strains harboring one of these mutations were resistant to more drugs than their competitors and displayed higher transmissibility. Therefore bacteria that acquired a compensatory mutation, cushioned the "cost" of drug resistance and were consequently able to accumulate even more drug resistances. It has been hypothesized that XDR strains are unable to infect people efficiently due to the fitness cost of their numerous resistances and that they do not constitute an epidemic threat (Burgos et al. 2003). Here we have demonstrated in natural epidemiologic settings that this paradigm does not hold for W148. Strains with compensatory mutations, despite higher number of resistances, can be more transmissible, confirming precedent predictions obtained via approximate Bayesian computation (Luciani et al. 2009). In a near future, inflating acquisition of compensatory mutations will pave the way for disastrous epidemics caused by MDR and XDR, and consequently pose serious public health issues. Moreover, our results highlight the urgent need to identify other, yet undetected, compensatory mutations and assess their presence while treating infected patients.

Interestingly, when analyzing differences between the basal clade and the W148 outbreak strains, we identified few novel antibiotic resistance SNPs, i.e. out of the scope of the common targets. Yet, three major characteristics must be featured.

First, at least 6 genes involved in immune response or epitope formation where affected by amino-acid replacements and up to 16 when including synonymous replacements, representing therefore up to 20% of all identified SNPs (Table 2). Discarding synonymous candidates can be misleading since there is increasing evidence for silent SNPs affecting antibiotic resistance (Safi et al. 2013). These

results highlight the major role of host pathogen interaction in the evolutionary drive of W148. What makes the results surprising is not the nature of the targeted genes, but rather the number of targets impacted. The split between the progenitor and the derived Russian strains is only 250 years old, but it has already led to a sharp antigenic set-up change. Although the reasons behind such a dramatic change remain speculative, one possibility could be the partial change of the host population (Li et al. 2008), shifting from an ancestral central Asian background to a derived/enriched Indo-European one (gulag prisoners and workers).

Second, two pairs of genes belonging to the type VII secretion system (T7SS) (Stanley et al. 2003; Abdallah et al. 2007) were affected by non-synonymous mutations. Rv3862c and Rv3871 are members of the ESX-1 secretion machinery, a secretion system especially responsible for virulence and macrophage escape (Hsu et al. 2003; Pym et al. 2003; McLaughlin et al. 2007). Upon the above-mentioned characteristics, the observed protein modifications can be therefore of major pathological significance. Two other genes were also involved, i.e. Rv3885c and Rv3887c, but those changes affected the less well-understood ESX-2 region.

Third, another specificity of the adaptive landscape of W148 is its sharp modulation in DNA damage response. Three RecA dependent genes (Rv3776, Rv2719c and Rv2579) involved in LexA binding regions showed amino acid changes. Moreover, they are likely binding to genes encoding small RNAs. This is exemplified by the presence of not less that 5 mutations on the sole Rv3776 gene. It is tempting to relate those features with elevated mutation and drug resistance rates in the Beijing lineage (de Steenwinkel et al. 2012; Ford et al. 2013a). Indeed, the SOS response, induced by DNA damage represents our best and most studied candidate for the genesis of transient mutators. Such phenotypes carry a short-term selective advantage owing to their capacity to generate adaptive mutations at a higher pace (Miller 1996), driving therefore their ubiquity in the bacteria phyla (Wirth et al. 2006).

Such virulence and adaptive up and down regulation mechanisms accompanied by genetic variability of the infecting bacillus are the scars of the selective pressures imposed by the host. These machineries share the same ontological goals, “tune and adapt”. To fully elucidate this hypothesis, extensive

research on *M. tuberculosis* antigens is needed to enhance our comprehension on immunity against tuberculosis and to facilitate the development of new vaccines and tuberculosis management tools (May et al. 2009). Last but not least, unraveling the mechanisms and pathways that modulate the mutation rate will accelerate the development of novel chemotherapy and potentially augment the activity of existing antibiotics.

Methods

Data collection and whole-genome sequencing

731 Beijing clade W148 strains were identified based on characteristic MTBC genotyping patterns or specific genetic polymorphisms, the global dataset entails 569 newly generated genomes, plus another 162 genomes retrieved from sequence read archives (see Table SX). Whole genome sequencing (WGS) was performed on all isolates using Illumina Technology (MiSeq and HiSeq 2500) with Nextera XT library preparation kits as instructed by the manufacturer (Illumina, San Diego, CA, USA). The raw data (fastq files) were submitted to the NCBI raw read archive (accession numbers pending).

Mapping of reads, SNP filtering and genotypic drug resistance prediction

Reads were mapped to the *M. tuberculosis* H37Rv genome (GenBank ID: NC_000962.3) using the exact alignment program SARUMAN (Blom et al. 2011). Polymorphisms with a minimum of 10x coverage and 75% variant frequency were extracted using customized perl scripts. After exclusion of drug resistance associated genes, repetitive regions and non-informative/non discriminating SNPs, the remaining positions that matched the above mentioned threshold levels in at least 95% of all isolates were considered as valid and used for all isolates in a concatenated sequence alignment.

To determine strains genotypic drug resistances we first considered all mutations in 28 coding sequences and four upstream regions associated with first- and second-line resistance (additional data). We excluded known phylogenetic variants for the resistance prediction according to Feuerriegel et al (Feuerriegel et al. 2012), as well as new combinations found to be monophyletic and linked to particular clades, e.g. Beijing clade W148, based on our phylogenetic reconstruction (additional data).

A mutation was considered as resistance associated in the context of this manuscript, when at least one of the following criteria was full-filled: associated with phenotypic resistance while no other possible resistance conferring mutation was found, i.e. all other related genes exhibit wild type alleles; included in commercial TB drug resistance test; found in mono resistant *in vitro* selected mutants (Nebenzahl-Guimaraes et al. 2014), identified as convergent mutation (different amino acid substitutions for the same codon), observed as homoplastic variant (found in different non-related phylogenetic subgroups), termed as high confident pyrazinamide resistance conferring mutation by Miotto et al (Miotto et al. 2014)

Likelihood mapping. The phylogenetic signal of the data set was investigated with the likelihood mapping method implemented in TREE -PUZZLE 7.1 (Schmidt et al. 2002) by analyzing 10,000 random quartets. This method proceeds by evaluating, using maximum-likelihood, groups of four randomly chosen sequences (quartets). The three possible unrooted tree topologies for each quartet are weighted, and the posterior weights are then plotted using triangular coordinates, such that each corner represents a fully resolved tree topology. Therefore, the resulting distribution of the points shows whether the data are suitable for a phylogenetic reconstruction.

Time-dependent phylogenetic and phylogeographic reconstruction

Phylogenetic relationships were reconstructed using the maximum-likelihood approach implemented in PHYML 3.412 (Guindon et al. 2010). The robustness of the maximum-likelihood tree topology was assessed with bootstrapping analyses of 1,000 pseudoreplicated data sets. Phylogenies were rooted with the midpoint rooting option using FIGTREE software v1.4 and with the reference *M. tuberculosis* strain H37Rv, both resulting in the same topology. The profile of drug resistances for each strains and information of compensatory mutations were plotted on the maximum likelihood tree using ITOL (Letunic and Bork 2016). Linear regression analysis of the root-to-tip distances against sampling time was performed using TEMPEST1.5 (Rambaut et al. 2016). To assess the robustness of our root-to-tip regression, we performed a permutation test of 5,000 replicates using the LMPERM Package (Anderson and Robinson 2001) in R. For the coalescent-based analyses, evolutionary rates and tree topologies were analyzed using the general time-reversible (GTR) and Hasegawa-Kishino-Yano (HKY)

substitution models with gamma distributed among-site rate variation with four rate categories (Γ4). The substitution rate was estimated under different demographic and clock models using BEAST v2.3.2 (Bouckaert et al. 2014) taking advantage of a sampling timeframe from 1995 to 2013. We tested both a strict molecular clock (which assumes the same evolutionary rates for all branches in the tree) and a relaxed clock that allows different rates among branches. Constant-sized and Bayesian skyline plot models, based on a general, non-parametric prior that enforces no particular demographic history were used. For each model, two independent chains were conducted for 100 million generations and convergence was assessed by checking ESS values for key parameters using TRACER V1.6. We used TRACER V1.6 to calculate the $\log_{10}$ bayes factors in order to compare the models after a burnin of 10% of the chain. Bayes factors represent the ratio of the marginal likelihood of the models being compared. Approximate marginal likelihoods for each coalescent model were calculated via importance sampling (1,000 bootstraps) using the harmonic mean of the sampled likelihoods. A ratio between 3 and 10 indicates moderate support that one model better fits the data than another, whereas values greater than 10 indicate strong support

W148 place of origin was detected with BEAST 2.3.2 (Bouckaert et al. 2014) and the GEO SPHERE package for a discrete trait phylogeography analysis in a Bayesian statistical framework (Lemey et al. 2009). As entries, for the strain's locations we used the country of origin of the patient if available, otherwise the country of isolation of the strain was retained. Moreover, to simplify the calculations we grouped the 24 countries represented in the dataset in 15 regions based on their number of isolates and area. We ran a chain of 300 million generations. Results were visualized on FIGTREE V1.4.2 and SPREAD. In addition, to compare regions of different sample sizes we calculated the nucleotide diversity *pi* per regions and performed bootstrap and subsampling of 19 individuals with 10,000 repetitions, in R, using the package APE (Paradis et al. 2004) and PEGAS (Paradis 2010). The minimum spanning tree was produced using BIONUMERICS version 7.6.

Statistical analysis and transmission index

To compare drug resistance differences between regions we performed chi-square tests and logistic regressions. To consider the association, positive or negative, between a drug resistance and a region, the two tests had to be significant ($P < 0.05$). Based on the distance matrix (SNP distances), we

further determined for every isolate the number of isolates that are in a range of 10 SNPs or less (in the following referred to as "transmission index"). Based on previous convergent estimates of MTBC genome evolution rate of ≈ 0.5 SNP/genome/year in inter-human transmission chains (Roetzer et al. 2013; Walker et al. 2013; Walker et al. 2014) and in macaque infection models (Ford et al. 2011), this 10 SNP-threshold was used to infer the number of recently (≤ 10 years) infected cases.

Impact of resistance-conferring and compensatory mutations on transmission success

We used multiple linear regression to examine the respective contributions of antimicrobial resistance and putative fitness cost-compensating mutations to the transmission success of tuberculosis. To take transmission duration into account, we computed, for each isolate and each period length T in years (from 1 to 40y before sampling), a transmission success score defined as the number of isolates distant of less than *T* SNPs, divided by *T*. This approach relied on the following rationale: based on MTBC evolution rate of 0.5 mutation per genome per year, the relation between evolution time and SNP divergence is such that a cluster with at most *N* SNPs of difference is expected to have evolved for approximately *N* years. Thus, transmission success score over T years could be interpreted as the size of the transmission network divided by its evolution time, hence as the average yearly increase of the network size. For each period *T*, the transmission success score was regressed on the number of resistance mutations and on the presence of putative compensatory mutations. The regression coefficients with 95% confidence intervals were computed and plotted against *T* to identify maxima, that is, time periods when the transmission success was maximally influenced by either resistance-conferring or - compensating mutations.

Detection of genes under positive selection

Additionally, to capture SNPs that may explain W148 success, we used the R version of the software PCADAPT (Duforet-Frebourg et al. 2014) to perform a genome scan based on a Bayesian factor model. We chose K = 4 factors and selected SNPs with the highest Mahalanobis distance and associated with the principal component 2, separating basal strains from W148 outbreak. The factor analysis was performed on the centered genotype matrix that was not scaled. The MCMC algorithm was initialized using singular value decomposition, and the total number of steps was equal to 400 with a burn-in of 200 steps.

Bioinformatics and protein structures

Protein functional information was obtained from TUBERCULIST (http://tuberculist.epfl.ch/). The mutated protein structures were predicted by PHYRE2 software online (http://www.sbg.bio.ic.ac.uk/phyre2) (Kelley et al. 2015).

Funding

Parts of the work were funded by the German Ministry of Education and Research (BMBF) for the German Center of Infection Research (DZIF).

Supplementary Figures and Tables

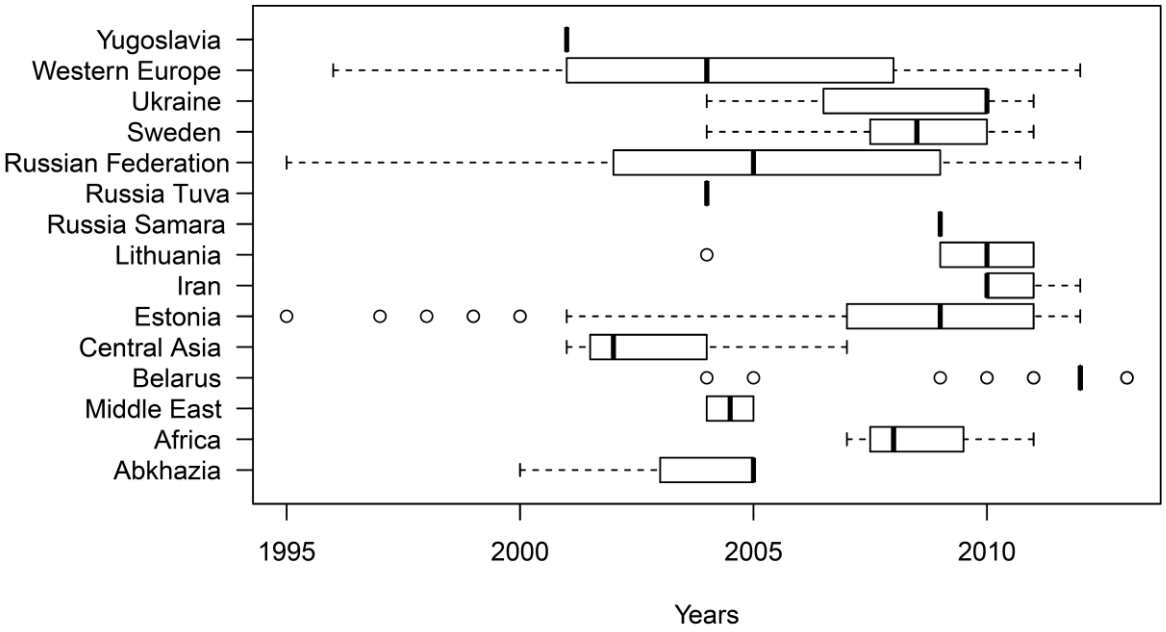


Fig. S1: Box-plots of the sampling years according to the regional boundaries.

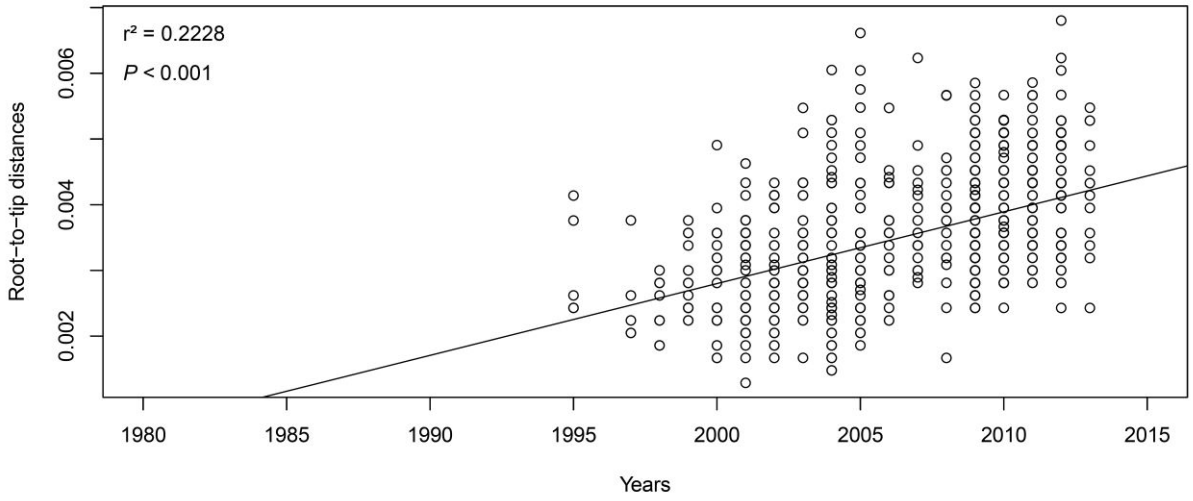


Fig. S2: Linear regression analysis showing correlation between root-to-tip distance and sampling years of the W148 strain collection covering the period 1995 to 2013.

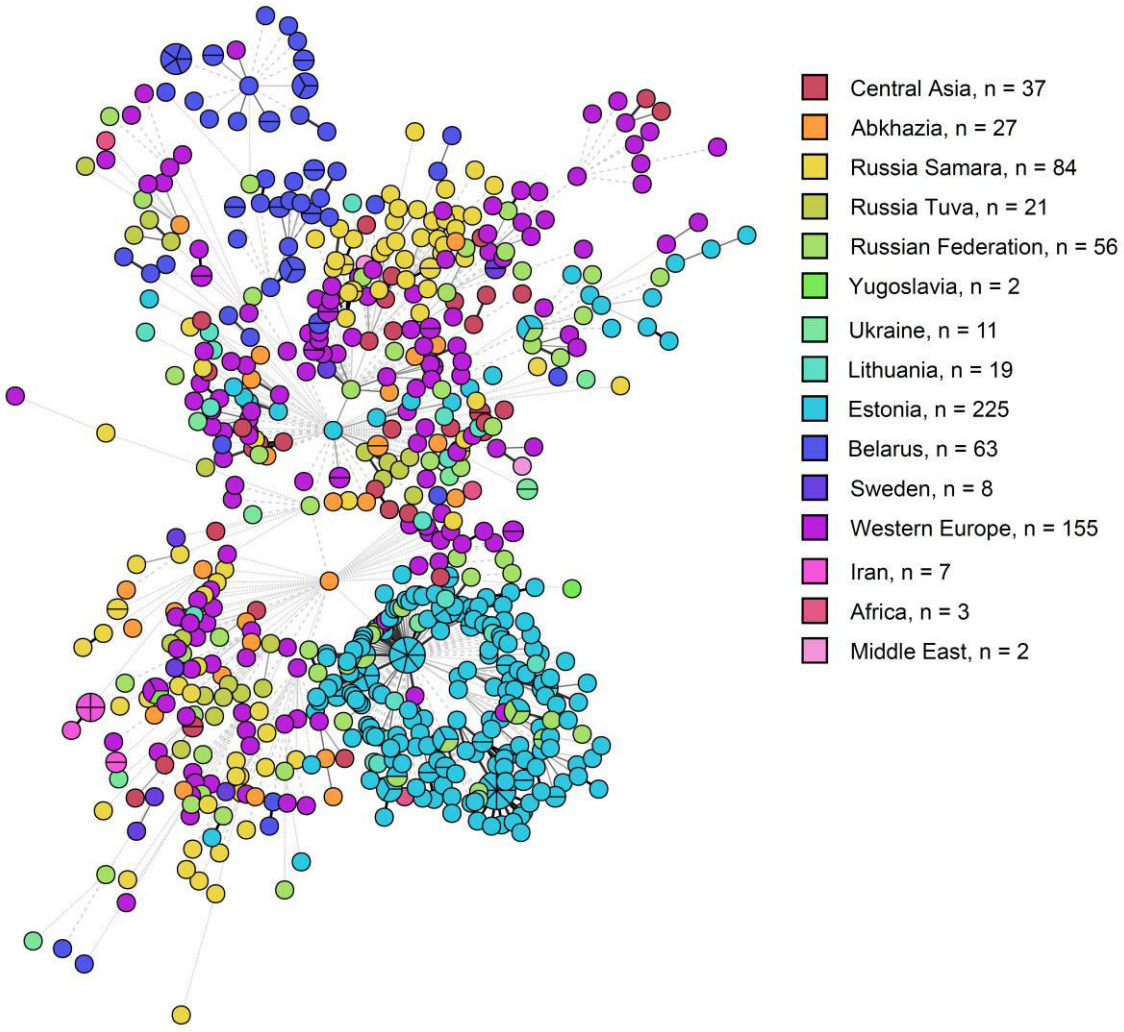


Fig. S3: Minimum spanning tree based on SNPs of the 720 W148 strains. Strains are colorized according to their location. The length of the links is proportional to the number of SNPs between strains.

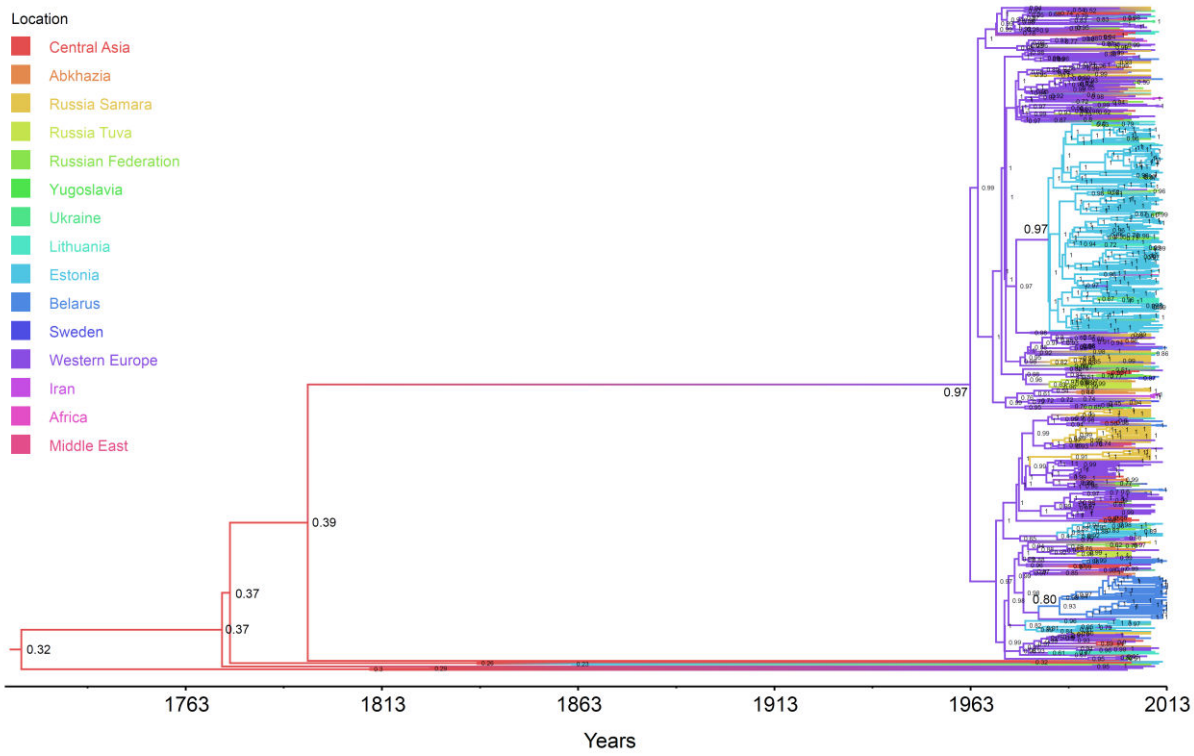


Fig. S4: W148 Bayesian maximum clade credibility phylogeny inferred under a skyline model using a HKY substitution model. Implementing a tip dating approach resulted in an estimated substitution rate of 1.12 x 10^{-7} substitutions per nucleotide per year. Branches are colored according to the most probable location state of their descendent node. The posterior probability of the most probable location is indicated above each branch.

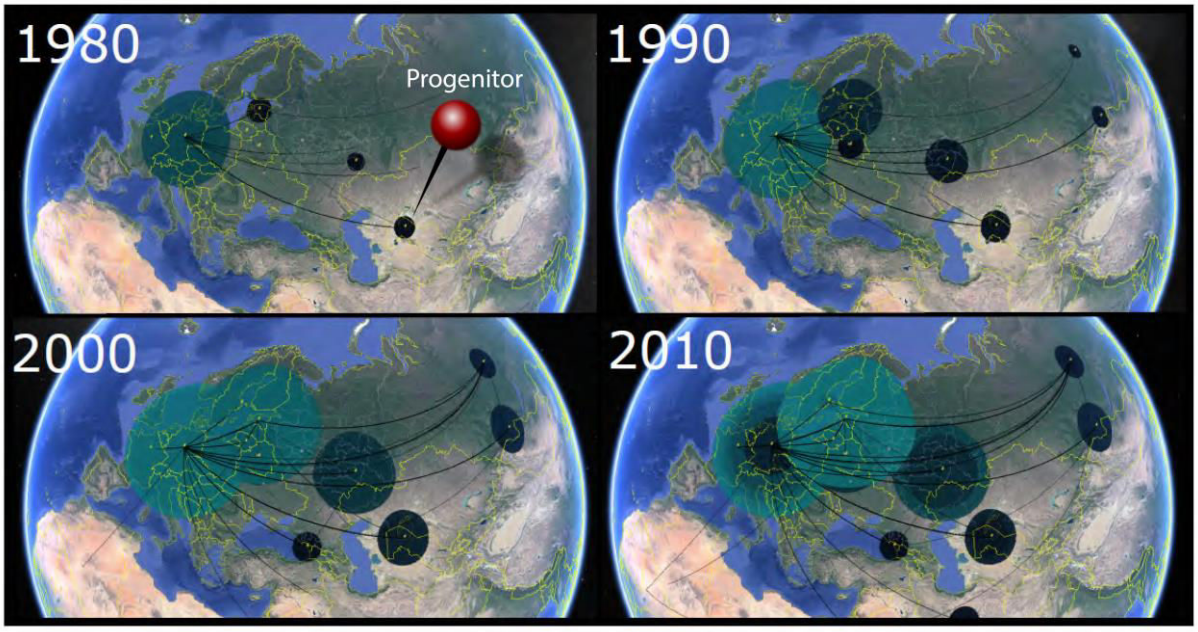


Fig. S5: W148 spread through time and space according to the Bayesian analyses.

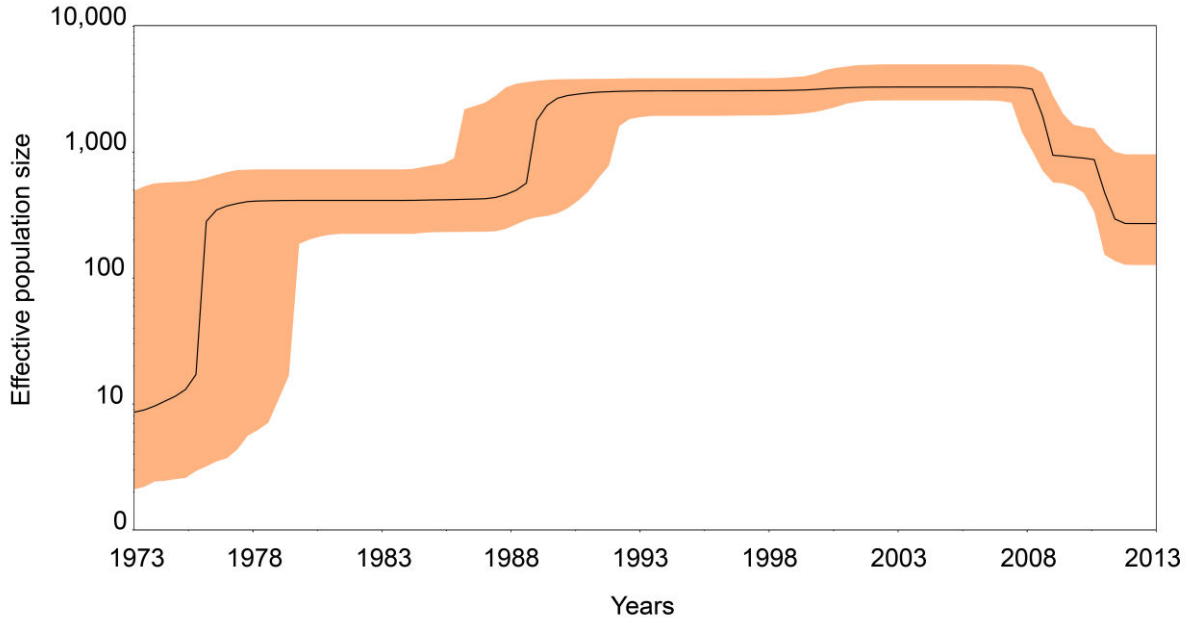


Fig. S6: Effective population size over time of the 720 W148 strains based on a Bayesian skyline approach using the HKY substitution model and a Log normal relaxed clock model estimated a mean mutation rate of 1.12 x 10^{-7} substitutions per nucleotide per year. Orange shaded area represents changes in the effective population size giving the 95% highest posterior density (HPD) interval with the black line representing the mean value.

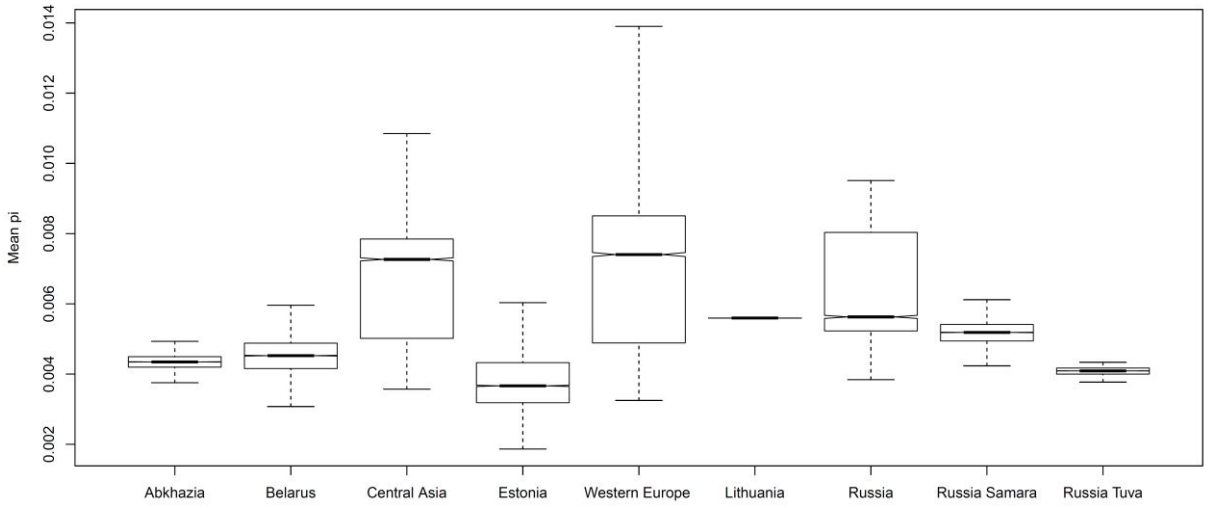


Fig. S7: Boxplot of the mean nucleotidic diversity calculated with subsampling of 19 strains of each region for 10,000 replicates (rarefaction procedure).

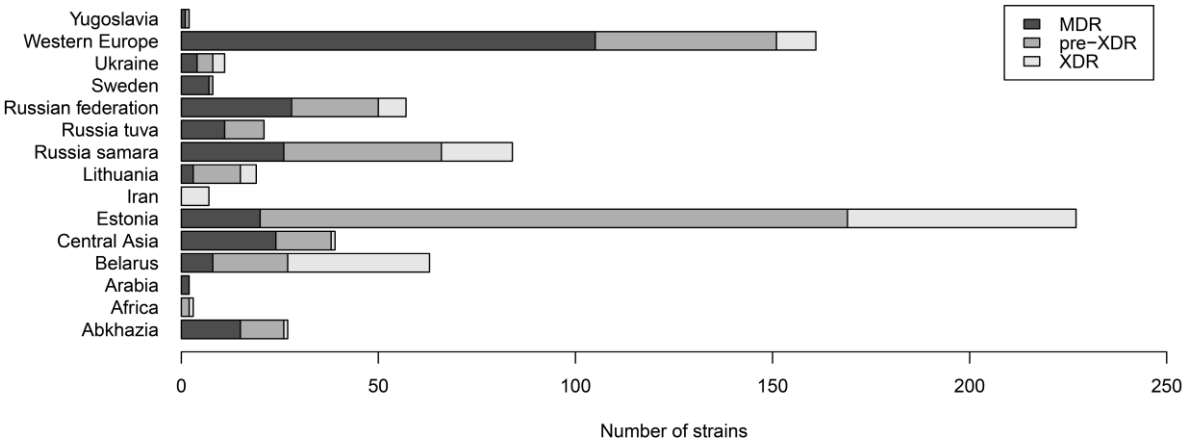


Fig. S8: Regional dependent antibiotic resistance profiles of the sample collection.

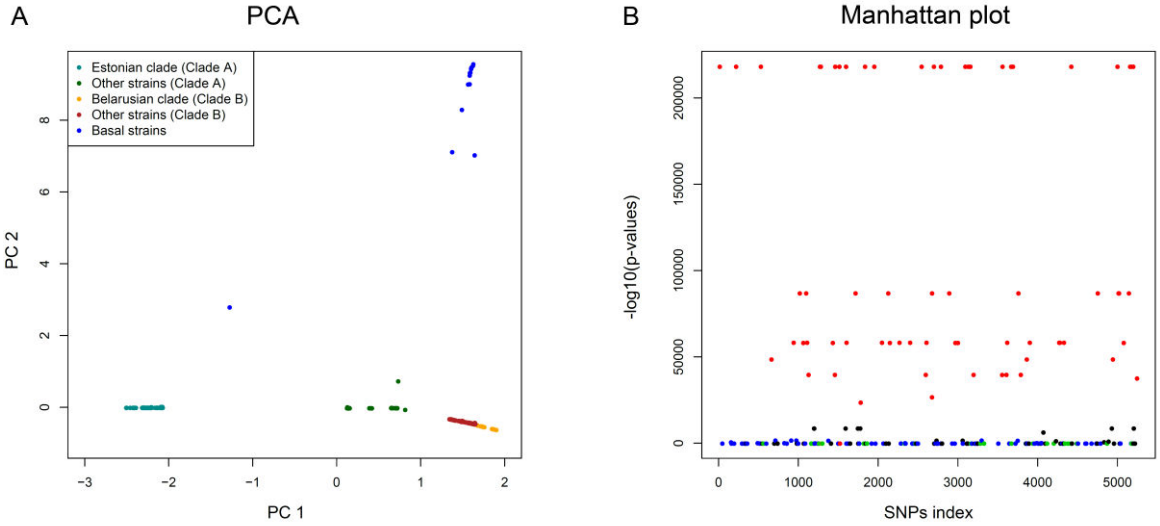


Fig. S9: Genome scan analysis for detecting SNPs involved in local adaptation. A) Plot of the first 2 principal components. The 731 strains are represented by points and colorized according to their phylogenic origin. The PC 2 is the one separating basal strains from others. B) Manhattan plot representing the 5264 SNPs and values obtained after performing Mahalanobis distances. The SNPs are colorized according to the pc to which they correlate most (PC1 = black, PC2 = red, PC3 = green and PC4 = blue).

References

Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CM, Appelmelk BJ, Bitter W. 2007. Type VII secretion--mycobacteria show the way. *Nature reviews* **5**(11): 883-891.

Anderson MJ, Robinson J. 2001. Permutation tests for linear models. *Australia and New Zealand Jounral of Statisitcs* **43**: 75-88.

Barbier M, Wirth T. 2016. The Evolutionary History, Demography, and Spread of the Mycobacterium tuberculosis Complex. *Microbiology spectrum* **4**(4).

Bespyatykh J, Shitikov E, Butenko I, Altukhov I, Alexeev D, Mokrousov I, Dogonadze M, Zhuravlev V, Yablonsky P, Ilina E et al. 2016. Proteome analysis of the Mycobacterium tuberculosis Beijing B0/W148 cluster. *Scientific reports* **6**: 28985.

Blom J, Jakobi T, Doppmeier D, Jaenicke S, Kalinowski J, Stoye J, Goesmann A. 2011. Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming. *Bioinformatics (Oxford, England)* **27**(10): 1351-1358.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS computational biology* **10**: 1-6.

Brandis G, Wrande M, Liljas L, Hughes D. 2012. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. *Molecular microbiology* **85**: 142-151.

Brodin P, Majlessi L, Marsollier L, de Jonge MI, Bottai D, Demangel C, Hinds J, Neyrolles O, Butcher PD, Leclerc C et al. 2006. Dissection of ESAT-6 system 1 of Mycobacterium tuberculosis and impact on immunogenicity and virulence. *Infection and immunity* **74**(1): 88-98.

Burgos M, DeRiemer K, Small PM, Hopewell PC, Daley CL. 2003. Effect of drug resistance on the generation of secondary cases of tuberculosis. *The Journal of infectious diseases* **188**: 1878-1884.

Chauhan A, Lofton H, Maloney E, Moore J, Fol M, Madiraju MV, Rajagopalan M. 2006. Interference of Mycobacterium tuberculosis cell division by Rv2719c, a cell wall hydrolase. *Molecular microbiology* **62**(1): 132-147.

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S. 2011. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature genetics* **44**: 106-110.

Davis EO, Dullaghan EM, Rand L. 2002. Definition of the mycobacterial SOS box and use to identify LexA-regulated genes in Mycobacterium tuberculosis. *Journal of bacteriology* **184**(12): 3287-3295.

de Steenwinkel JE, ten Kate MT, de Knegt GJ, Kremer K, Aarnoutse RE, Boeree MJ, Verbrugh HA, van Soolingen D, Bakker-Woudenberg IA. 2012. Drug susceptibility of Mycobacterium tuberculosis Beijing genotype and association with MDR TB. *Emerging infectious diseases* **18**(4): 660-663.

De Vos M, Müller B, Borrell S, Black PA, Van Helden PD, Warren RM, Gagneux S, Victor TC. 2013. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrobial agents and chemotherapy* **57**: 827-832.

Duforet-Frebourg N, Bazin E, Blum MGB. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular biology and evolution* **31**: 1-13.

Dullaghan EM, Brooks PC, Davis EO. 2002. The role of multiple SOS boxes upstream of the Mycobacterium tuberculosis lexA gene--identification of a novel DNA-damage-inducible gene. *Microbiology (Reading, England)* **148**(Pt 11): 3609-3615.

Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. 2015. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature communications* **6**: 7119.

Eldholm V, Pettersson JH-O, Brynildsrud OB, Kitchen A, Rasmussen EM, Lillebaek T, Rønning JO, Crudu V, Mengshoel AT, Debech N et al. 2016. Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*: 201611283.

Faustini A, Hall AJ, Perucci CA. 2006. Risk factors for multidrug resistant tuberculosis in Europe: a systematic review. *Thorax* **61**: 158-163.

Ferraris DM, Spallek R, Oehlmann W, Singh M, Rizzi M. 2014. Crystal structure of the Mycobacterium tuberculosis phosphate binding protein PstS3. *Proteins* **82**(9): 2268-2274.

Feuerriegel S, Oberhauser B, George AG, Dafae F, Richter E, Rüsch-Gerdes S, Niemann S. 2012. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. *BMC microbiology* **12**: 90.

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M et al. 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature genetics* **43**(5): 482-486.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013a. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics* **45**(7): 784-790.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013b. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics* **45**: 784-790.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC et al. 2006. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 2869-2873.

Gagneux S, Small PM. 2007. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infectious Diseases* **7**: 328-337.

Gao LY, Guo S, McLaughlin B, Morisaki H, Engel JN, Brown EJ. 2004. A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion. *Molecular microbiology* **53**(6): 1677-1693.

Gerasimova A, Kazakov AE, Arkin AP, Dubchak I, Gelfand MS. 2011. Comparative genomics of the dormancy regulons in mycobacteria. *Journal of bacteriology* **193**(14): 3446-3452.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Mehtods to Estimate Maximum-Likelihood Phylogenies: Asessing the Performance of PhyML 2.0. *Systematic biology* **59**: 307-321.

Handel A, Regoes RR, Antia R. 2006. The role of compensatory mutations in the emergence of drug resistance. *PLoS computational biology* **2**: 1262-1270.

Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ, Morin PM, Marks CB, Padiyar J, Goulding C, Gingery M et al. 2003. The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proceedings of the National Academy of Sciences of the United States of America* **100**(21): 12420-12425.

Iseman MD. 2002. Tuberculosis therapy: past, present and future. *European Respiratory Journal* **20**: 87S-94s.

Kassa D, Ran L, Geberemeskel W, Tebeje M, Alemu A, Selase A, Tegbaru B, Franken KL, Friggen AH, van Meijgaarden KE et al. 2012. Analysis of immune responses against a wide range of Mycobacterium tuberculosis antigens in patients with active pulmonary tuberculosis. *Clinical and vaccine immunology : CVI* **19**(12): 1907-1915.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**(6): 845-858.

Lemey P, Rambaut A, Drummond AJ, Suchard Ma. 2009. Bayesian phylogeography finds its roots. *PLoS computational biology* **5**.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* **44**: W242-W245.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, NY* **319**(5866): 1100-1104.

Li QJ, Jiao WW, Yin QQ, Xu F, Li JQ, Sun L, Xiao J, Li YJ, Mokrousov I, Huang HR et al. 2016. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant Mycobacterium tuberculosis Beijing genotype strains in China. *Antimicrobial agents and chemotherapy* **60**: 2807-2812.

Liu L, Zhang WJ, Zheng J, Fu H, Chen Q, Zhang Z, Chen X, Zhou B, Feng L, Liu H et al. 2014. Exploration of novel cellular and serological antigen biomarkers in the ORFeome of Mycobacterium tuberculosis. *Molecular & cellular proteomics : MCP* **13**(3): 897-906.

Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM. 2009. The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 14711-14715.

May YL, Reddy TBK, Arend SM, Friggen AH, Franken KLMC, Van Meijgaarden KE, Verduyn MJC, Schoolnik GK, Klein MR, Ottenhoff THM. 2009. Cross-reactive immunity to Mycobacterium tuberculosis DosR regulon-encoded antigens in individuals infected with environmental, nontuberculous mycobacteria. *Infection and immunity* **77**: 5071-5079.

McLaughlin B, Chon JS, MacGurn JA, Carlsson F, Cheng TL, Cox JS, Brown EJ. 2007. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. *PLoS pathogens* **3**(8): e105.

Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, Rüsch-Gerdes S, Mokrousov I, Aleksic E et al. 2015. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage (Supplementary). *Nature genetics*.

Miller JH. 1996. Spontaneous mutators in bacteria: insights into pathways of mutagenesis and repair. *Annual review of microbiology* **50**: 625-643.

Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, Bakonyte D, Stakenas P, Pimkina E, Augustynowicz-Kopeć E et al. 2014. Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *mBio* **5**: e01819-01814.

Mokrousov I. 2013. Insights into the origin, emergence, and current spread of a successful Russian clone of Mycobacterium tuberculosis. *Clinical microbiology reviews* **26**: 342-360.

Nebenzahl-Guimaraes H, Jacobson KR, Farhat MR, Murray MB. 2014. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis. *The Journal of antimicrobial chemotherapy* **69**: 331-342.

Nei M, Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**: 5269-5273.

Nieselt-Struwe K, von Haeseler A. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular biology and evolution* **18**(7): 1204-1219.

Paradis E. 2010. Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics (Oxford, England)* **26**: 419-420.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics (Oxford, England)* **20**: 289-290.

Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC, Glynn JR, Crampin AC, Alves A et al. 2016. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC genomics* **17**: 151.

Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, Williams A, Griffiths KE, Marchal G, Leclerc C, Cole ST. 2003. Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nature medicine* **9**(5): 533-539.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* **2**: vew007.

Renton AM, Borisenko KK, Meheus A, Gromyko A. 1998. Epidemics of syphilis in the newly independent states of the former Soviet Union. *Sex Transm Infect* **74**: 165-166.

Ribeiro SC, Gomes LL, Amaral EP, Andrade MR, Almeida FM, Rezende AL, Lanes VR, Carvalho EC, Suffys PN, Mokrousov I et al. 2014. Mycobacterium tuberculosis strains of the modern sublineage of the Beijing family are more likely to display increased virulence than strains of the ancient sublineage. *Journal of clinical microbiology* **52**: 2615-2624.

Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rusch-Gerdes S et al. 2013. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS medicine* **10**(2): e1001387.

Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, McNeil M, Peterson SN, Chatterjee D, Fleischmann R et al. 2013. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45**(10): 1190-1197.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics (Oxford, England)* **18**(3): 502-504.

Serra-Vidal MM, Latorre I, Franken KL, Diaz J, de Souza-Galvao ML, Casas I, Maldonado J, Mila C, Solsona J, Jimenez-Fuentes MA et al. 2014. Immunogenicity of 60 novel latency-related antigens of Mycobacterium tuberculosis. *Frontiers in microbiology* **5**: 517.

Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrell MI, Schoolnik GK. 2001. Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. *Proceedings of the National Academy of Sciences of the United States of America* **98**(13): 7534-7539.

Shitikov EA, Bespyatykh JA, Ischenko DS, Alexeev DG, Karpova IY, Kostryukova ES, Isaeva YD, Nosova EY, Mokrousov IV, Vyazovaya AA et al. 2014. Unusual large-scale chromosomal rearrangements in Mycobacterium tuberculosis Beijing B0/W148 cluster isolates. *PLoS ONE* **9**: 1-9.

Skrahina A, Hurevich H, Zalutskaya A, Sahalchyk E, Astrauko A, Hoffner S, Rusovich V, Dadu A, de Colombani P, Dara M et al. 2013. Multidrug-resistant tuberculosis in Belarus: the size of the problem and associated risk factors. *Bulletin of the World Health Organization* **91**: 36-45.

Smollett KL, Smith KM, Kahramanoglou C, Arnvig KB, Buxton RS, Davis EO. 2012. Global analysis of the regulon of the transcriptional repressor LexA, a key component of SOS response in Mycobacterium tuberculosis. *The Journal of biological chemistry* **287**(26): 22004-22014.

Stanley SA, Raghavan S, Hwang WW, Cox JS. 2003. Acute infection and macrophage subversion by Mycobacterium tuberculosis require a specialized secretion system. *Proceedings of the National Academy of Sciences of the United States of America* **100**(22): 13001-13006.

van der Werf MJ, Zellweger JP. 2016. Impact of migration on tuberculosis epidemiology and control in the EU/EEA. *Eurosurveillance* **21**: 8-11.

Vitek CR, Wharton M. 1998. Diphtheria in the former Soviet Union: Reemergence of a pandemic disease. *Emerging infectious diseases* **4**: 539-550.

Vyazovaya A, Mokrousov I, Solovieva N, Mushkin A, Manicheva O, Vishnevsky B, Zhuravlev V, Narvskayaa O. 2015. Tuberculous spondylitis in Russia and prominent role of multidrug-resistant clone Mycobacterium tuberculosis Beijing B0/W148. *Antimicrobial agents and chemotherapy* **59**: 2349-2357.

Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW et al. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases* **13**: 137-146.

Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP et al. 2014. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *The lancet Respiratory medicine* **2**: 285-292.

WHO. 2011. Estonia Tuberculosis country work summary. 2010-2011.

WHO. 2016. Global Tuberculosis Report 2016. *Cdc 2016*: 214.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H et al. 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. *Molecular microbiology* **60**(5): 1136-1151.

Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S et al. 2008. Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS pathogens* **4**(9): e1000160.

Zvi A, Ariel N, Fulkerson J, Sadoff JC, Shafferman A. 2008. Whole genome identification of Mycobacterium tuberculosis vaccine candidates by comprehensive data mining and bioinformatic analyses. *BMC medical genomics* **1**: 18.

Chapitre 4.

Changements d'hôtes au sein du complexe *Mycobacterium tuberculosis* et leurs conséquences adaptatives

Chapitre 4 : Changements d'hôtes au sein du complexe *Mycobacterium tuberculosis* et leurs conséquences adaptatives

Introduction

Le MTBC (*Mycobacterium tuberculosis* complex) représente toutes les « espèces » proches de *M. tuberculosis,* l'agent étiologique de la tuberculose chez l'homme. Toutes les lignées le composant ne sont distantes génétiquement que de 0.05% au maximum (Rodriguez-Campos et al. 2014) mais pourtant infectent toute une variété de mammifères. On peut citer entre autre *M. bovis* infectant les bovins (Karlson 1970), *M. caprae* touchant les caprins (Aranaz et al. 2003), *M. origys* pathogène des oryx (van Ingen et al. 2012), *M. suricattae* comme son nom l'indique touche les suricates (Dippenaar et al. 2015), *M. mungi* chez les mangoustes (Alexander et al. 2010). Les rongeurs sont infectés par *M. microti* (Boniotti et al. 2014) et les pinnipèdes (phoques, otaries) par *M. pinnipedii* (Cousins et al. 2003). Récemment une nouvelle lignée a également été isolé chez un chimpanzé (Coscolla et al. 2013). Il est remarquable de constater que toutes les espèces d'hôtes concernées sont des espèces sociables vivant en groupes d'individus, tout comme l'homme. Cette panoplie d'espèces « candidates » vient renforcer la nature épidémique de *M. tuberculosis*, pathogène aérotransmissible, qui rencontre un setting de transmission favorable dans des populations à hautes densités.

La bactérie la plus proche du MTBC est un microorganisme environnemental, pouvant être pathogène opportuniste, nommée *M. canettii* (Van Soolingen et al. 1997). Ses caractéristiques morphologiques diffèrent des membres du MTBC, elle forme des colonies lisses tandis que les membres du MTBC forment des colonies à l'apparence rugueuse. Comparativement à *M. tuberculosis*, *M. canettii* semble pousser plus rapidement, être moins virulente, persister un temps inférieur dans son hôte mais également ne pas se transmettre d'hôte à hôte (Supply et al. 2013). Uniquement identifier chez des patients, son potentiel réservoir est inconnu mais son origine environnementale est fortement suggérée au vu de différents type de CRISPR-Cas présent sur son génome et fortement apparentés avec ceux de bactéries environnementales (milieux halophiles et extrêmophiles). La diversité génétique spécifique à *M. canettii* est bien supérieure à celle du MTBC, et contrairement à ce dernier, l'analyse du génome des différentes souches semblent indiquer des traces importantes de recombinaison (Supply et al. 2013). Face à tous ces indices, l'ancêtre du MTBC est fortement soupçonné d'avoir les propriétés des souches *M. canettii* actuelles, et d'être devenu un pathogène exclusif suite à l'infection opportuniste d'un homme. Suite à ces conclusions, il semble donc naturel

d'enraciner la phylogénie globale du MTBC avec une souche de *M. canettii,* comme illustré dans plusieurs études récentes (Bos et al. 2014; Comas et al. 2013).

La tuberculose a longtemps été vue comme une zoonose, passant des bovins à l'homme lors de la domestication (Smith et al. 2009). Le séquençage de *M. bovis* et l'analyse de régions de délétions ont permis de montrer que l'inverse avait vraisemblablement eu lieu, et que la tuberculose serait possiblement passée de l'homme à l'animal (Brosch et al. 2002). De plus, dans une étude utilisant les MIRUs, il a été montré que le MTBC pouvait être divisé en deux clades, l'un purement humains et le second contenant des lignées humaines (lignées 5 et 6) et les différentes lignées animales, confirmée plus tard par WGS (Wirth et al. 2008; Bos et al. 2014). Les lignées 5 et 6 sont des lignées humaines confinées à l'Afrique de l'Ouest qui semblent moins virulentes que les souches des autres lignées humaines. La cause de cette baisse de pathogénicité par rapport aux souches humaines pourrait être une érosion de leur génome, de la même manière que chez les souches animales, et notamment des mutations au sein de gènes impliqués dans la virulence (Gonzalo-Asensio et al. 2014; Bentley et al. 2012). Le génome des souches animales est souvent vu comme déliquescent, entrainant une perte de la capacité à être transmis entre hommes pour ces organismes qui ne provoquent que de rares épisodes infectieux (Berg and Smith 2014). Cependant de la même manière, *M. tuberculosis* ne se transmet pas d'animaux à animaux et les cas d'hommes infectant des animaux sont rares. Plutôt que des pathogènes ayant perdu une partie de leur pouvoir de virulence chez l'homme, ces souches devraient plutôt être étudiées comme des bactéries s'étant totalement adapté à leur hôte principal (Behr and Gordon 2015), au même titre que *M. tuberculosis* s'est adapté à l'homme. Cependant un retour en arrière ne semble pas impossible, des souches de *M. pinnipedii* ont été identifiées comme ayant provoqué des infections chez 3 Amérindiens il y a environ 1000 ans, au niveau de l'actuel Pérou (Bos et al. 2014). Les auteurs ont proposé l'hypothèse que la tuberculose a pu être amenée en Amérique, avant la découverte du nouveau monde par les européens, par des pinnipèdes qui l'auraient transmise aux populations locales.

Afin d'étudier l'adaptation des différentes espèces du MTBC à leurs hôtes respectifs nous avons collecté des souches représentant toutes les lignées de tuberculose, rassemblant un total de 308 organismes en y incluant le maximum de souches animales disponibles dans les bases de données. Par ailleurs, nous avons ajouté les génomes complets de 40 souches animales séquencés par nos soins. Ceci est l'une des premières études génomiques couvrant une aussi grande diversité de souches animales.

Méthodes

Séquençage et SNP calling

Pour cette étude, nous avons fait séquencer par MiSeq (technologie Illumina) en génomes complets 24 souches de *M. pinnipedii*, 24 de *M.* bovis et 11 de *M. microti.* Les autres séquences sont celles utilisées dans l'article de Bos *et al.* 2014. Les fichiers fastq pour le *Chimpanzee bacillus* étaient si gros que nous avons sous échantillonné 3 fois 1,200,000 séquences différentes afin d'obtenir 3 répliquas pour un même organisme.
Toutes les séquences obtenues ont été assemblées par mapping en utilisant H37rv comme souche de référence grâce à l'algorithme BWA dans BIONUMERICS 7.6. Toujours avec ce logiciel, nous avec effectué le SNP calling en utilisant comme paramètres un distance inter-SNPs supérieure ou égale à 10 bases, une couverture de minimum 5x par bases avec au moins 1 lecture forward et 1 reverse. De plus nous avons éliminé toutes les positions contenant au moins une délétion ou une base ambiguë. Ces paramètres stringents nous ont permis d'extraire et de concaténer des SNPs de haute confiance. Le SNP calling a été effectué à deux reprises, une première fois avec l'ensemble des génomes séquencés et seconde fois uniquement avec les souches des lignées 5, 6 et animales afin de produire deux jeux de données.

Reconstructions phylogénétiques

Nous avons produit deux types de phylogénies dans cette étude, des phylogénies en Maximum Likelihood avec le logiciel JMODELTEST v2.1 et déterminé le modèle de substitution optimal par BIC (Bayesian Information Criterion). Et des phylogénies bayésiennes dans BEAST (Drummond and Ho 2007). Nous avons respectivement utilisé FIGTREE et DENSITREE pour représenter les arbres.

Recherche de SNPs sous sélection

Afin de chercher et d'identifier des SNPs possiblement sous sélection et ayant joué un rôle dans l'adaptation des souches animales à leur hôte nous avons établi un protocole en deux étapes. La 1ère consiste à identifier des sites soit sous sélection diversifiante soit fixées dans les souches associées à des lignées animales. Pour cette première étape nous avons donc, d'une part cherché des SNPs homoplastiques et donc possiblement sous sélection positive. Pour cela nous avons calculé le score d'homoplasie de chaque SNPs, par rapport aux arbres produits en Maximum Likelihood précédemment, dans PAUP (Swofford 2002). D'autre part

nous avons recherché les SNPs possiblement en lien avec une adaptation locale à l'aide du logiciel PCADAPT (Duforet-Frebourg et al. 2015). Dans les deux cas, une liste des mutations et des gènes concernés sera dressée. Dans un second temps, les gènes identifiés à la première étape seront analysés avec CODEML du package PAML (Yang 2007). Cet algorithme permet de détecter avec une plus grande finesse les sites sous sélection positive ($\omega > 1$) sur les protéines en accompagnant cette information d'un support statistique (likelihood).

Résultats et discussions

Après assemblage des séquences, nous avons effectué deux filtrages différents. Le premier pour obtenir le jeu de données global, composé de 310 séquences de 26898 SNPs dont la composition peut être observée en table 1, et le second, comprenant uniquement les lignées 5 et 6 ainsi que les souches de la lignée animale représentant un total de 89 séquences d'une longueur de 13109 SNPs.

Lignée	Nombre de génomes
Lineage 1	38
Lineage 2	69
Lineage 3	33
Lineage 4	54
Lineage 5	14
Lineage 6	14
Lineage 7	27
Mycobacterium bovis	18
Mycobacterium caprae	4
Mycobacterium pinnipedii	26
Mycobacterium microti	8
Mycobacterium suricattae	1
Mycobacterium orygis	1
Chimpanzee bacillus	3

Table 1 : Nombre de séquences par lignées humaines et animales inclues dans l'étude.

Nous avons produit un arbre phylogénétique du jeu de donnée global par Maximum Likelihood (Figure 1). Nous n'avons pas utilisé de souche *M. canettii* pour l'enracinement car comme évoqué plus haut, cette espèce bactérienne recombine énormément et est relativement éloignée du MTBC, entrainant possiblement des distorsions lors de la reconstruction de

l'arbre (branches longues, effets confondant des HGT). Son emplacement au niveau de la phylogénie nous semblant trop incertain, nous avons préféré ne pas l'inclure et utiliser en enracinement par midpoint rooting, stratégie qui paraît solide au vu de la forte clonalité du pathogène et de l'accumulation régulière des mutations au fil du temps. La topologie de l'arbre obtenu est congruente avec celle d'études antérieures, plus particulièrement en ce qui concerne les lignées purement humaines 1, 7, 4, 3 et 2 regroupées ensemble. Une certaine différence apparaît cependant lorsque l'on scrute les lignées 5 et 6, ainsi que les lignées animales. En effet, la topologie de l'arbre indique que la lignée 5 est un groupe frère proche des lignée humaines modernes. La lignée 6 et les souches animales formant un clade « profond » distinct. Il est intéressant de constater que la souche *M. suricattae* et le *Chimpanzee bacillus,* dont les 3 répliquas n'ont pas un SNP de différence entre eux, groupent avec la lignée 6 plutôt qu'avec les autres lignées animales. *M. orygis, M. caprae et M. bovis* peuvent être regroupés et se séparent du clade *M. pinnipedii et M. microti*. En prenant uniquement le jeu de données comprenant les lignées 5 et 6 et les lignées animales, nous obtenons la même topologie, cette fois obtenue par inférence bayésienne (Figure 2). Les branches du densitree obtenues sont très claires, indiquant un bon support et une probabilité postérieure élevée. Cela montre qu'il y a peu ou pas de points de conflit dans la topologie de cet arbre.

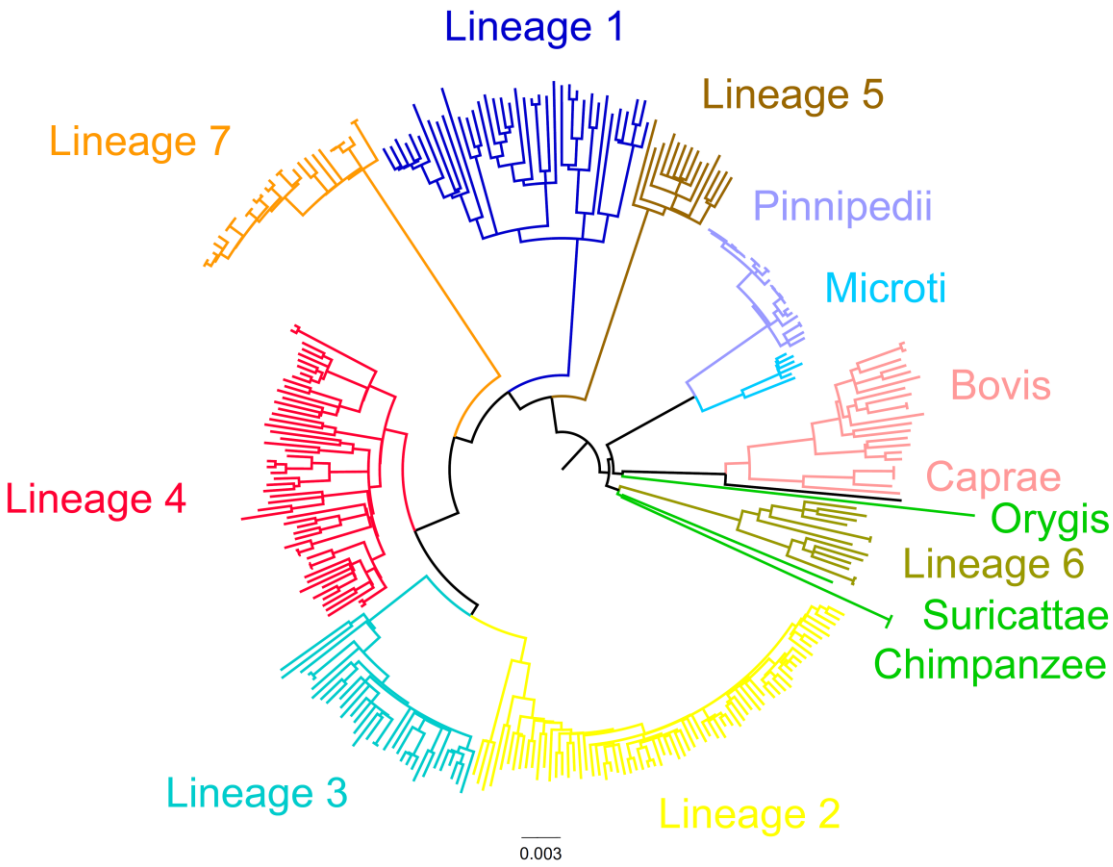


Figure 1 : Arbre représentatif du MTBC obtenu par Maximum Likelihood enraciné en midpoint rooting. Nous avons utilisé le logiciel JMODELTEST et le BIC afin de déterminer le modèle de substitution à utiliser. Le modèle de substitutions obtenant le score de BIC le plus élevé était le modèle TVM.

Au vu du fait que la lignée 6 ait un ancêtre commun avec *M. suricattae* et *Chimpanzee bacillus* plus récent que la séparation d'avec les autres lignées animales, on est en droit de s'interroger sur l'hôte de la souche ancestrale. S'agit-il d'un pathogène humain ou animal ? Il est possible que plus tôt dans l'évolution du MTBC, cette souche ne présentait pas encore d'hôte préférentiel et ce n'est qu'avec le temps et l'isolation que de vraies lignées animales adaptées à leur hôte sont apparues. A ce jour *M. canettii* n'a été identifié que chez des patients humains, mais jamais chez des animaux. Si *M. canettii* pouvait infecter également des animaux, étant normalement proche de la souche ancestrale du MTBC, alors cela renforcerait l'hypothèse d'une souche ancestrale non spécialisée. Une autre alternative peut correspondre à la situation observée sur l'origine de la malaria, où une origine des souches humaines provenant de grands primates se fait de plus en plus vraisemblable (Loy et al. 2017). Ce scénario nécessite le séquençage d'autres génomes de Mycobactéries pulmonaires provenant par exemple de gorilles.

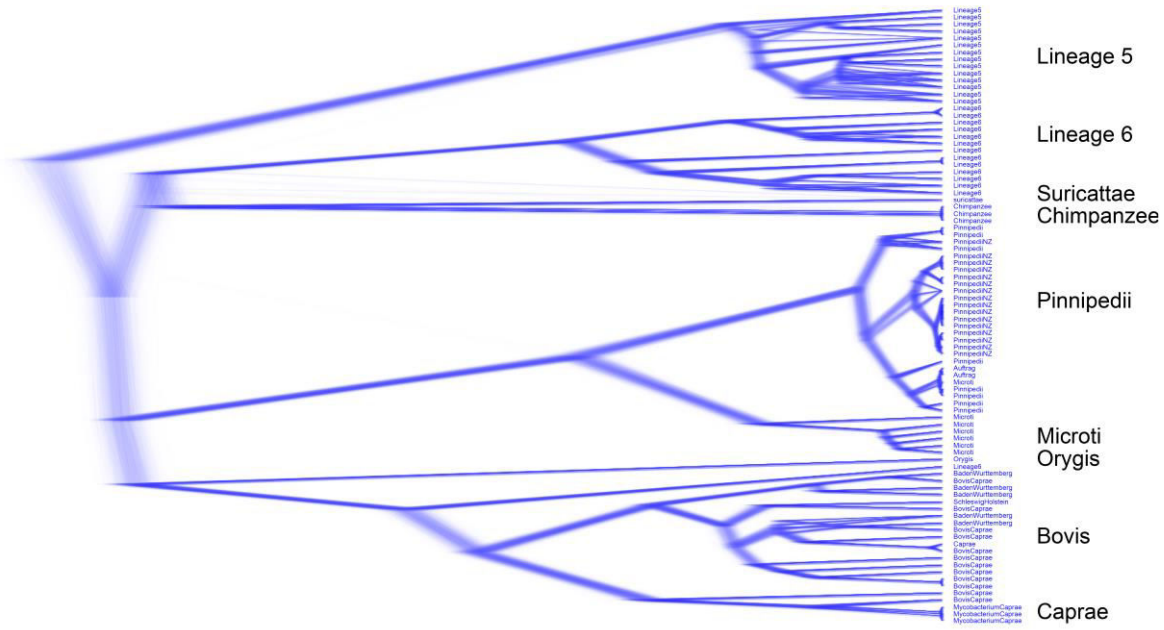


Figure 2 : DensiTree des souches animales et des lignées 5 et 6. Les arbres ont été inférés en utilisant BEAST. Nous avons sélectionné le modèle de substitution HKY, avec un taux de mutation constant.

Nous avons utilisé le logiciel PCADAPT afin d'identifier des gènes possiblement impliqués dans l'adaptation des souches à différents hôtes. On peut voir que l'axe 1 de l'ACP différencie la lignée 6 de la lignée 5 et des souches animales. L'axe 2 par contre différencie les souches *M. pinnipedii* des souches *M. bovis.* (Figure 3). En observant le Manhattan plot, les SNPs présentant les Bayes factors les plus élevés sont les SNPs associés avec l'axe 2. Une liste des 20 premiers SNPs détectés est disponible en Table 2. Nous devons encore déterminer la position génomique correspondant aux SNPs et le gène associé. Logiquement ces mutations devraient être fixées de façon variable au sein des souches *M. bovis* ou *M. pinnipedii* et sont donc possiblement impliquées dans l'adaptation de la mycobactérie aux pinnipèdes pour certaines et aux bovins pour les autres.

Les analyses de cette étude sont encore en cours et tous les résultats ne sont donc pas disponibles. Ce chapitre est encore à l'état d'ébauche et j'en ai bien conscience. Nous discuterons la suite de cet article et les analyses envisagées dans la conclusion.

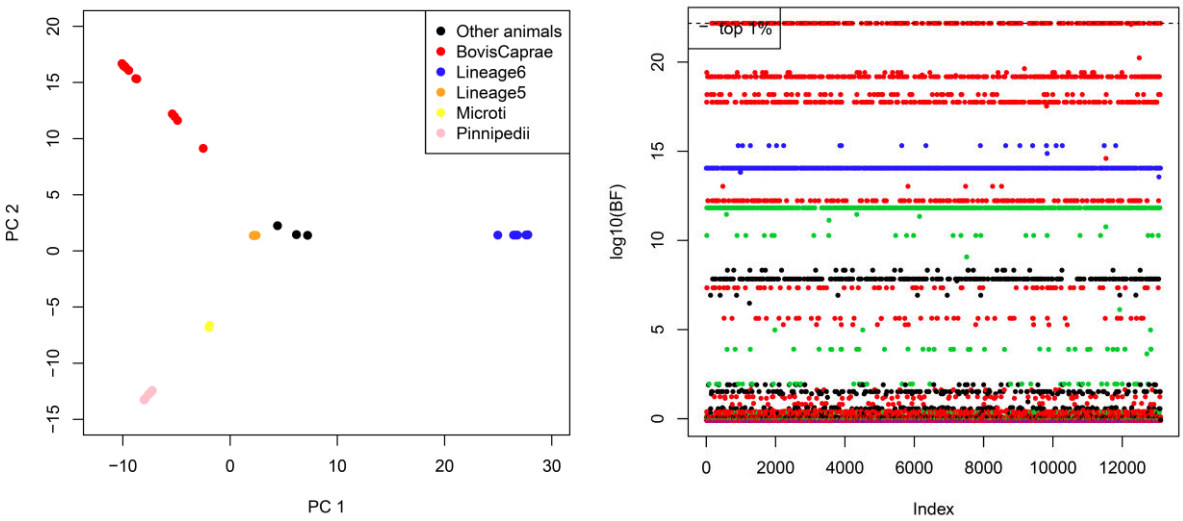


Figure 3 : Résultats de l'analyse PCAdapt. A gauche on peut observer les deux premiers axes de l'ACP. Le premier axe sépare la lignée 6 des autres tandis que le second axe sépare les souches *M. pinnipedii* des souches *M. bovis* et *M. caprae*. A droite on peut observer Manhattan plot inférés pour les SNPs. Plus le Bayes factor d'un SNP est élevé et plus celui-ci a une probabilité élevée d'être associé à une adaptation locale. La couleur du SNP indique l'axe avec lequel il est associé. En noir le PC1, en rouge le PC2, en vert le PC3 et en bleu le PC4.

SNPs	Factor	BF
158	2	1.49239E+22
176	2	1.49221E+22
504	2	1.49212E+22
576	2	1.4921E+22
578	2	1.4921E+22
627	2	1.49223E+22
757	2	1.49218E+22
899	2	1.4922E+22
917	2	1.49234E+22
1056	2	1.49265E+22
1156	2	1.49224E+22
1336	2	1.49236E+22
1531	2	1.49241E+22
1580	2	1.49222E+22
1635	2	1.49217E+22
1744	2	1.49241E+22
1818	2	1.49207E+22
1842	2	1.49228E+22
1875	2	1.49221E+22
1909	2	1.49229E+22

Table 2 : Liste des 20 premiers SNPs identifiés lors de l'analyse PCAdapt.

Bibliographie

Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, van Pittius NCG. 2010. Novel Mycobacterium tuberculosis complex pathogen, M. Mungi. *Emerg Infect Dis* **16**: 1296–1299.

Aranaz A, Cousins D, Mateos A, Domínguez L. 2003. Elevation of Mycobacterium tuberculosis subsp. caprae Aranaz et al. 1999 to species rank as Mycobacterium caprae comb. nov., sp. nov. *Int J Syst Evol Microbiol* **53**: 1785–1789.

Behr M a, Gordon S V. 2015. Why doesn ' t Mycobacterium tuberculosis spread in animals ? *Trends Microbiol* **23**: 1–2.

Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, Thurston S, Gagneux S, Wood J, Antonio M, et al. 2012. The genome of mycobacterium Africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. *PLoS Negl Trop Dis* **6**.

Berg S, Smith NH. 2014. Why doesn't bovine tuberculosis transmit between humans? *Trends Microbiol* **22**: 552–553.

Boniotti MB, Gaffuri A, Gelmetti D, Tagliabue S, Chiari M, Spisani M, Nassuato C, Gibelli L, Sacchi C, Zanoni M, et al. 2014. Detection and molecular characterization of Mycobacterium microti in wild boar from northern Italy. *J Clin Microbiol* **52**: 2834–2843.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest S a., Bryant JM, Harris SR, Schuenemann VJ, et al. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*.

Brosch R, Gordon S V, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, et al. 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci U S A* **99**: 3684–3689.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**: 1176–1182.

Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, Dabrowski PW, Radonic A, Niemann S, Parkhill J, et al. 2013. Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. *Emerg Infect Dis* **19**: 969–976.

Cousins D V., Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, Duignan P, Murray A, Dupont C, Ahmed N, et al. 2003. Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *Int J Syst Evol Microbiol* **53**: 1305–1314.

Dippenaar A, Parsons SDC, Sampson SL, van der Merwe RG, Drewe JA, Abdallah AM, Siame KK, Gey van Pittius NC, van Helden PD, Pain A, et al. 2015. Whole genome sequence analysis of Mycobacterium suricattae. *Tuberculosis* **95**: 682–688.

Drummond AJ, Ho SYW. 2007. Manual BEAST 1.4. *Edinburgh …* 1–41.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. 2015. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *Mbe* 1504.04543.

Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, Laval F, Daffé M, Martin C, Brosch R, et al. 2014. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A* **111**: 11491–11496.

Karlson AG. 1970. Mycobacterium Bovis Nom. Nov. *Int J Syst Bacteriol* **20**: 273–282.

Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH. 2017. Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int J Parasitol* **47**: 87–97.

Rodriguez-Campos S, Smith NH, Boniotti MB, Aranaz A. 2014. Overview and phylogeny of Mycobacterium tuberculosis complex organisms: Implications for diagnostics and legislation of bovine tuberculosis. *Res Vet Sci* **97**: S5–S19.

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon S V. 2009. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol* **7**: 537–544.

Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, et al. 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nat Genet* **45**: 172–9.

Swofford DL. 2002. Phylogenetic Analysis Using Parsimony. *Options* **42**: 294–307.

van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, van Soolingen D. 2012. Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies. *Emerg Infect Dis* **18**: 653–655.

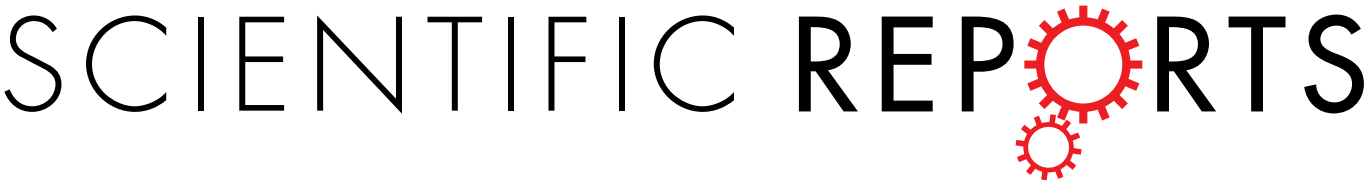
Van Soolingen D, Hoogenboezem T, De Haas PEW, Hermans PWM, Koedam MA, Teppema KS, Brennan PJ, Besra GS, Portaels F, Top J, et al. 1997. A Novel Pathogenic Taxon of the Mycobacterium tuberculosis Complex, Canetti: Characterization of an Exceptional Isolate from Africa. *Int J Syst Bacteriol* **47**: 1236–1245.

Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, Van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S, et al. 2008. Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog* **4**.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Chapitre 5.

L’estimation souche spécifique du succès épidémique contribue à la compréhension des dynamiques de transmission au sein de la tuberculose



OPEN

Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis

Received: 31 October 2016
Accepted: 21 February 2017
Published: 28 March 2017

Jean-Philippe Rasigade[1,2,3,4], Maxime Barbier[1,2], Oana Dumitrescu[3,4], Catherine Pichat[4], Gérard Carret[4], Anne-Sophie Ronnaux-Baron[5], Ghislaine Blasquez[5], Christine Godin-Benhaim[6], Sandrine Boisset[7,8], Anne Carricajo[9], Véronique Jacomo[10], Isabelle Fredenucci[4], Michèle Pérouse de Montclos[4], Jean-Pierre Flandrois[4,11], Florence Ader[3,12], Philip Supply[13], Gérard Lina[3,4] & Thierry Wirth[1,2]

The transmission dynamics of tuberculosis involves complex interactions of socio-economic and, possibly, microbiological factors. We describe an analytical framework to infer factors of epidemic success based on the joint analysis of epidemiological, clinical and pathogen genetic data. We derive isolate-specific, genetic distance-based estimates of epidemic success, and we represent success-related time-dependent concepts, namely epidemicity and endemicity, by restricting analysis to specific time scales. The method is applied to analyze a surveillance-based cohort of 1,641 tuberculosis patients with minisatellite-based isolate genotypes. Known predictors of isolate endemicity (older age, native status) and epidemicity (younger age, sputum smear positivity) were identified with high confidence ($P < 0.001$). Long-term epidemic success also correlated with the ability of Euro-American and Beijing MTBC lineages to cause active pulmonary infection, independent of patient age and country of origin. Our results demonstrate how important insights into the transmission dynamics of tuberculosis can be gained from active surveillance data.

The tuberculosis (TB) agent *Mycobacterium tuberculosis* complex (MTBC) has plagued mankind for millennia and, in spite of important efforts to slow down its progression, will probably continue to do so for decades[1,2]. TB prevalence is highly contrasted between world regions. Most patients with TB live in low-income countries while prevalence can be very low in high-income countries. Such a prevalence contrast, along with increasing population movements and migrations, has led to a situation in which the TB epidemiology and the MTBC population structure in low-prevalence areas is nowadays strongly impacted by influx of TB patients originating from high-prevalence areas[3–6]. Even single events of exogenous strain introduction in a low prevalence area can lead to rapid epidemic spread and large TB transmission clusters after a certain period of time in certain contexts[7,8].

[1]Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Université Pierre et Marie Curie, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France. [2]Laboratoire Biologie Intégrative des Populations, Ecole Pratique des Hautes Etudes, PSL Research University, Paris, France. [3]Centre International de Recherche en Infectiologie, CIRI, University of Lyon, France. [4]Institut des Agents Infectieux, Hospices Civils de Lyon, Lyon, France. [5]Comité Départemental d'Hygiène Sociale, CLAT69, Lyon, France. [6]Agence Régionale de Santé Rhône-Alpes, Lyon, France. [7]Laboratoire de Bactériologie, Institut de Biologie et de Pathologie, CHU de Grenoble, Grenoble, France. [8]Laboratoire TIMC-IMAG, UMR 5525 CNRS-UJF, UFR de Médecine, Université Grenoble Alpes, Grenoble, France. [9]Laboratoire des Agents Infectieux et d'Hygiène, CHU de Saint-Etienne, Saint-Etienne, France. [10]Laboratoire Biomnis, Lyon, France. [11]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, University of Lyon, France. [12]Service des Maladies Infectieuses et Tropicales, Hôpital de la Croix-Rousse, Hospices Civils de Lyon, Lyon, France. [13]INSERM U1019, CNRS-UMR 8204, Center for Infection and Immunity of Lille, Institut Pasteur de Lille, Université de Lille, Lille, France. Correspondence and requests for materials should be addressed to J.P.R. (email: jean-philippe.rasigade@univ-lyon1.fr) or T.W. (email: wirth@mnhn.fr)

Having the ability to capture the transmission dynamics and the epidemic success over time of particular strain groups from contemporary bacterial populations, and to identify associated contributions of pathogen- and/or host-related factors, could thus have important implications for epidemiological control and the understanding of bacterial evolution. In principle, past population dynamics of pathogens and the contribution of pathogen- or host-associated factors could be inferred from studies combining bacterial genetic data with patient clinical or socio-demographic data. Indeed, inferences based on population genetics methods and the coalescent theory, such as the skyline plot estimates of the evolution of population size over time[9–11], have been successfully used by our group[12,13] and others[14–16] to detect important demographic events in MTBC history such as, for instance, episodes of strong expansion of the Beijing MTBC lineage during the Industrial Revolution and the First World War. However, current coalescent-based methods analyze correlates of epidemic success at broad strain group levels, such as species or lineages, rather than on individual strains[17]. Therefore, these methods inherently carry the risk of mixing strains with distinct demographic histories, potentially averaging out important strain-specific characteristics. Conversely, performing separate analyses on smaller groups of isolates substantially increases the uncertainty of the demographic estimates[18].

In this work, we postulated that proxy measures of bacterial population dynamics such as epidemic success, endemicity and epidemicity, can be estimated at the level of each individual isolate in a study population. After demonstrating the relevance of this approach in simulations, we investigated a diversified MTBC population, typical of those seen in low TB prevalence areas[4], obtained from a cohort of 1,641 TB patients from the Rhône-Alpes region of France. Our analysis discriminated isolates of epidemic strain groups introduced recently in the region from those of the regional endemic background. Finally, the inclusion of isolate-level estimates of epidemic success in regression-based association analyses identified both expected and novel links between MTBC transmission dynamics and the characteristics of patient and strain groups in our setting.

Results

Estimating epidemic success from genetic distances. Proposing a quantitative correlate of the epidemic success of a pathogen is difficult owing to the lack of a formal and consensual definition of epidemic success[19]. Here we define epidemic success as a purely quantitative and time-dependent concept: the epidemic success of a bacterial group is proportional to the frequency of its associated transmission events during a given period of time.

All else equal, and assuming a strain transmission rate that is higher than strain mutation rate (which is reasonable for TB)[20], epidemic success in a successful group increases prevalence faster than diversity, resulting in a more clonal (i.e., less diverse) structure compared to other groups in the sample. Lower diversity results into smaller genetic distances between isolates. From a statistical standpoint, both the prevalence of, and pairwise genetic distances between isolates in a group can be jointly quantified by a measure of density in the space of genetic distances, suggesting that density correlates with success. Importantly, density is defined for all points in the space of genetic distances, hence on the level of individuals in the population. Based on this rationale, we postulated that a measure of density associated with the haplotype of an isolate reflects the epidemic success of its ancestors compared to other isolates in the sampled population.

We constructed the density measure using an application-specific adaptation of a classical non-parametric technique, namely kernel density estimation (KDE)[21]. In the general case, KDE computes density based on distances between points and a kernel function, endowed with a bandwidth parameter to control the smoothness of the estimate. In our application, points were haplotypes, distances were the pairwise numbers of allelic differences and the kernel function was based on the geometric distribution. To control the bandwidth of the analysis in an interpretable fashion, we expressed this bandwith as a timescale parameter equal to the median time to the most recent common ancestor (TMRCA50, see Methods) under the kernel distribution and an evolutionary rate known a priori (Fig. 1). Intuitively, the timescale allows one to focus the analysis on recent transmission events (e.g. to detect epidemic isolates with short-term success) or to extend this focus towards the past (e.g. to detect endemic isolates with long-term success). In the following, we refer to KDE-based density estimates as timescaled haplotypic densities (THDs).

Timescaled haplotypic density correlates with epidemic success in silico. To investigate how THD reflects expansion events (epidemic bursts), we generated synthetic sets of haplotypes by means of Fastsimcoal 2 software[22]. Model parameters were carefully selected to mimic MTBC populations with genotypes obtained from independent minisatellite loci (as is the case in our cohort), with an evolutionary rate $\mu = 5 \times 10^{-4}$ change per locus per year, selected as the average of previous estimates[13,23–26] ranging from $\mu = 10^{-4}$ to 10^{-3}; a generation time of one day; and a contemporary effective population size $N_0 = 10^7$ as determined from our previous analysis of MTBC haplotypes obtained from minisatellite data[13]. To simulate the success of a pathogen population, we used a scenario in which independent epidemic subpopulations emerge from a constant-size ($N_0 = 10^7$) basal population and grow exponentially during 100y to reach the same contemporary population size as the basal population. Simulations used expansion factors up to 500-fold over 100y (~6% yearly increase), of the same order of magnitude as previous estimates of the expansion of successful MTBC clades[13,15,27]. Scaled geometric means of THDs per population (see Methods) with a 20y timescale and varying sample sizes, numbers of VNTR loci and fold-change expansions of the epidemic subpopulations are shown in Fig. 2. Additional simulations using 1 and 10 kbp DNA sequences in place of VNTRs are depicted in Supplementary Fig. S1. Collectively, these results demonstrate that: i) THD correlates with population expansion; ii) expectedly, estimation accuracy increases with sample size and, to a lesser extent, with the number of genetic loci; and iii) scaled THDs are invariant relative to the number of markers.

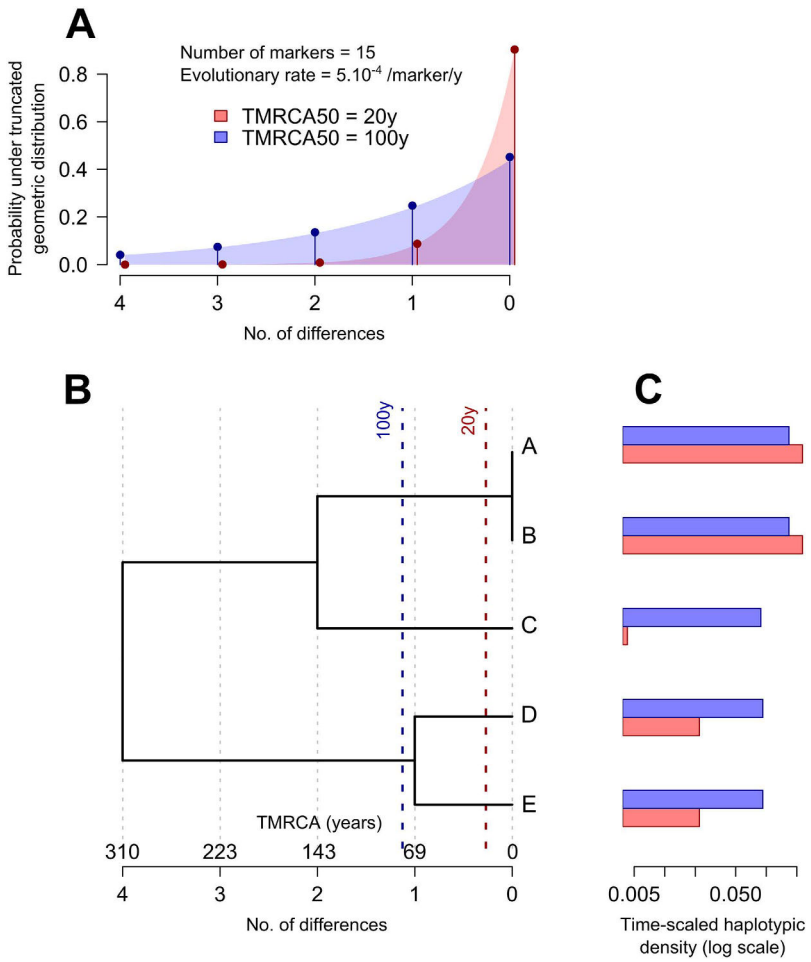


Figure 1. Time-scaled haplotypic density. THD computations were exemplified using a synthetic set of five 15-marker haplotypes (panel B). The timescales were defined as the median of a geometric distribution expressed in units of time (dashed lines in panel B; time units indicated above the X axis), based on the functional relationship between the genetic distance and the time to the most recent common ancestor (TMRCA; see Methods). Pairwise genetic distances were then associated with probabilities under the truncated geometric distribution (panel A). Probabilities decreased with the distance in a timescale-dependent fashion, with a faster decrease using the shorter 20y timescale (red curve) compared to the 100y timescale (blue curve). For each haplotype, THD was defined as the average of the probabilities associated with the distance from this haplotype to the others (panel C). Using a short timescale, haplotypes A and B, which have close relatives in the population, had much larger THDs compared to haplotype C, which has no close relative (red bars). Using a longer timescale, haplotype C had THD similar to that of haplotypes D and E because the densities of their respective clades were comparable relative to the timescale (blue bars). Remark the larger variance of the THD estimates with a short timescale compared to the larger timescale.

Characteristics of MTBC-infected patients in the Rhône-Alpes region of France. We investigated a collection of MTBC isolates representative of the Rhône-Alpes region of France, a low-MTBC prevalence area[28]. A total of 1,641 unique MTBC isolates (i.e. all from different patients) were recovered from the database of the Observatoire Rhône-Alpin des Mycobactéries (ORAM), a regional network of healthcare institutions involved in tuberculosis diagnosis and surveillance, from 2008 to 2014. Based on surveillance data available for the year 2010, our cohort included approximately 55% of all newly diagnosed TB patients in the region (see Supplementary Methods). Available socio-demographic, clinical and microbiological data, including indications of the proportion of missing data, are summarized in Table 1. French-native patients accounted for one-third of cases, consistent with previous reports in similar low-prevalence settings[5]. Rates of multidrug-resistance and resistance to first-line antibiotics rifampicin and isoniazid were 3.7, 3.8 and 10.1%, respectively.

MTBC population structure. Two classical complementary genotyping methods were performed on the 1,641 MTBC isolates included, namely spoligotyping[29] and MIRU-VNTR typing with a standard 15-locus scheme[30]. Spoligotyping is based on the detection of a collection of unique spacer sequences in a CRISPR locus[29,31]. Spoligotypes can be compared to databases to assign the strain to a family, a sublineage or a lineage. MIRU-VNTR typing interrogates multiple genomic loci containing variable numbers of tandem repeats.

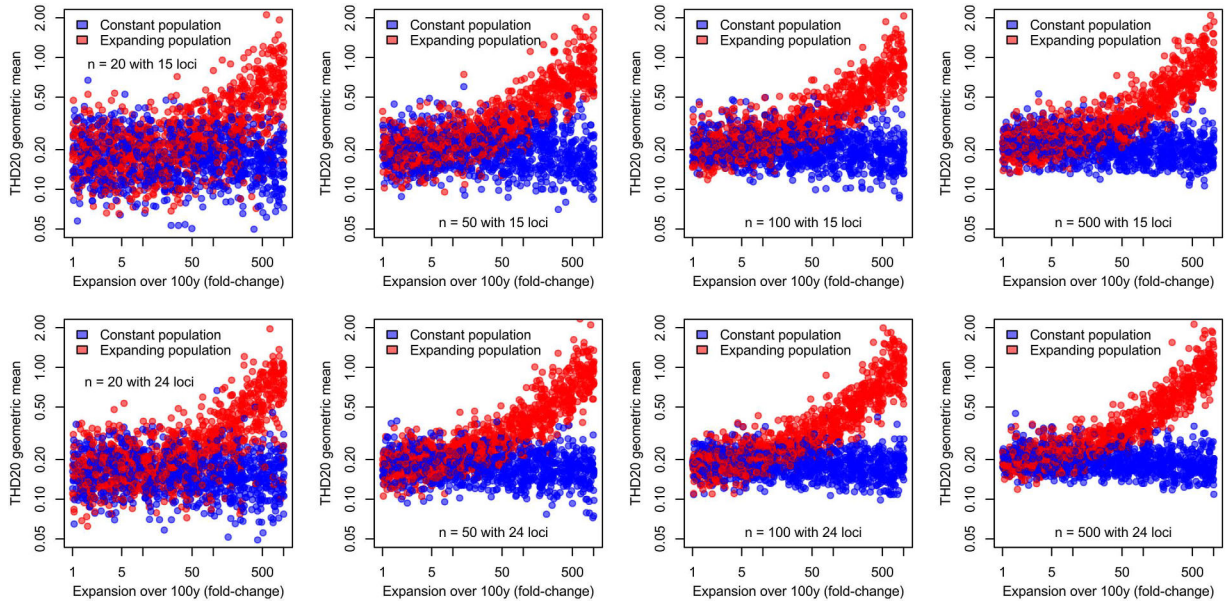


Figure 2. Timescaled haplotypic density (THD) of simulated constant-size and expanding populations. Markers represent scaled THD geometric means for 1,000 simulated metapopulations per panel, each comprising of a basal population with constant effective size (blue) and an epidemic population expanding with exponential growth over 100y (red) with varying expansion fold-change (X-axis), sample size per population and number of genetic loci.

Factor[a]		Cases with available data (%)
Median age at diagnosis [IQR]	48 [31–72]	1636 (99.7)
Male sex (%)	947 (57.7)	1640 (99.9)
French-native (%)	225 (32.9)	683 (41.6)
Median time in France before diagnosis for non-native patients [IQR]	5 [0–15][b]	156 (36.4)
Collective dwelling (%)	65 (24.3)	268 (16.3)
Occupation (%)	—	227 (13.8)
Employed	69 (30.4)	—
Retired	76 (33.5)	—
Student	36 (15.9)	—
Unemployed	46 (20.3)	—
Pulmonary infection (%)	656 (71.2)	921 (56.1)
AFB-positive sputum (%)	288 (44.4)	648 (39.5)
Rifampicin resistance (%)	32 (3.8)	836 (50.1)
Isoniazid resistance (%)	69 (10.1)	679 (41.4)
Multidrug resistance (%)	25 (3.7)	679 (41.4)

Table 1. Socio-demographic and disease-related characteristics of 1,641 MTBC-infected patients from the French Rhône-Alpes region, 2008–2014. [a]Numbers of patients or strains (%) unless specified otherwise; IQR, interquartile range.

Compared to spoligotyping, the resolution power of MIRU-VNTR typing for distinguishing MTBC strains is higher, and as such, it can be used as a proxy for inferring recent transmission of MTBC strains[12]. MIRU-VNTR typing is also more robust-although clearly imperfect compared to DNA sequences, due to homoplasy-than spoligotyping for phylogenetic classification[31], and it has been used successfully to investigate population dynamics at the level of MTBC lineages[13,32–34].

Spoligotypes were compared to those of SpolDB4 database[35], which allowed us to assign isolates into families including AFRI, Beijing, BOV, Cameroon, CAS, Haarlem, LAM, S, T and X[31], which were then reclassified into 6 major genome sequence- (or genomic deletion-) based lineages, including e.g. the East-African Indian, East Asian, Euro-American, Indo-Oceanic and West African lineages[36], according to known correspondences[37]. Strains of *M. bovis*, *M. pinnipedii* and *M. microti* were assigned to the so-called Animal lineage on a same

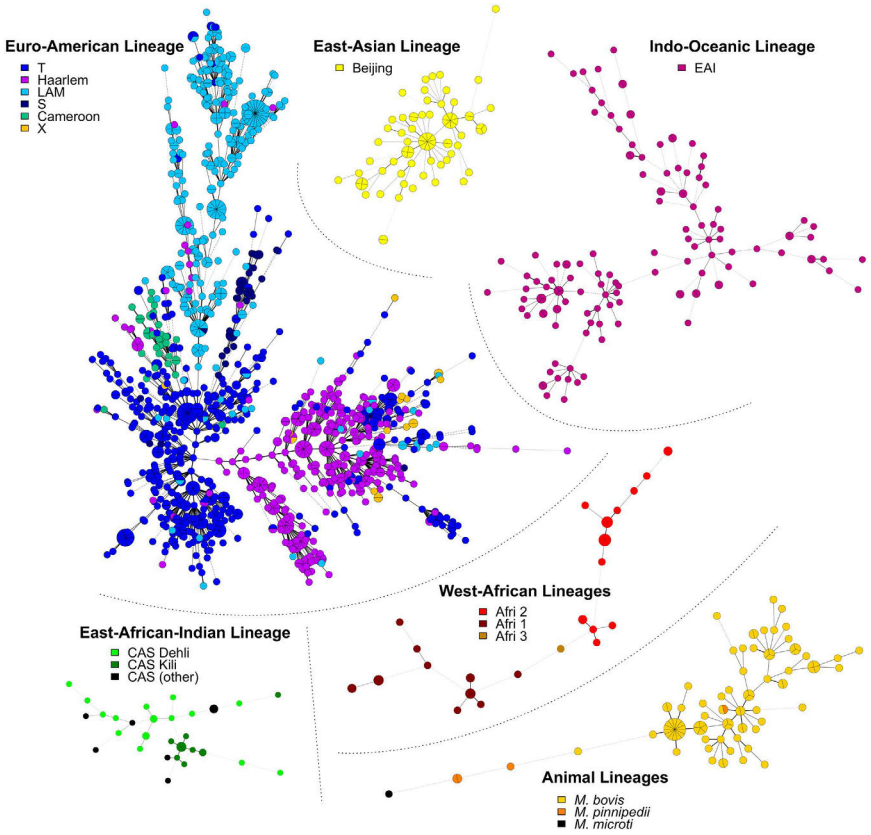


Figure 3. Population structure of MTBC strains isolated from the Rhône-Alpes region of France. Shown are independent MSTrees (one per major lineage) based on 15-loci MIRU-VNTR typing of 1,641 MTBC strains isolated from 2008 to 2014. Lengths of links between nodes are proportional to the number of allelic differences. Larger graph nodes indicate clusters of strains with identical MIRU genotypes. Node colors indicate spoligotype-based families.

basis. The resulting groupings and correspondences between families (such as Haarlem) and lineages (such as Euro-American) are made apparent in Fig. 3. Strains of the Euro-American lineage were most prevalent (see Supplementary Table S1).

Minimum spanning trees (MSTrees) were constructed based on the 15-loci MIRU-VNTR haplotypes to obtain graphical representations of the relationships between MIRU-VNTR haplotypes within each lineage (Fig. 3). We then investigated how THD analyses correlated with MSTree structures to illustrate how the qualitative and subjective information provided by MStrees is captured by THD in a quantitative and objective fashion. Short- and long-term THD timescales of 20 and 200y, respectively, were used in our analyses. The 200y timescale approximately matched the onset of the Industrial Revolution, previously reported to coincide with the expansion of several MTBC lineages[12,13]. The 20y timescale was chosen arbitrarily to reflect transmission over a much shorter period of time, of the same order of magnitude as a human generation. Additionally, 20y can be considered the shortest informative timescale with respect to MIRU-VNTR evolution (using an evolutionary rate of 5×10^{-4} change per locus per year, the probability of observing a change among 15 independent markers over 20y is ~14%).

To separately investigate diversity and prevalence of lineages and families, THDs were computed either relative to the complete strain collection (hereafter, global THDs) or to each lineage or family, independently (herafter, within-group THDs). In both cases, log-THDs were normalized and summarized as means and 95% confidence intervals of the mean (Fig. 4). Comparisons of global THDs allowed describing the evolutionary success of each group relative to the other groups, taking both prevalence and genetic diversity into account. Within-group THDs ignored the global population structure, mostly reflecting clonality in each group independent of their prevalence or genetic relatedness with other groups. Detailed insights into the relationships of spoligotype family or lineage, timescale and THD measures are provided in Supplementary Fig. S2.

Within-lineage and -family THDs, shown as red markers in Fig. 4, reflected the structural characteristics of the MSTrees inferred from the same groups, shown in Fig. 3. The highest long-term within-lineage THD was found in East-Asian/Beijing strains, consistent with the dense, radial structure of their respective MSTree, suggestive of recent population expansion and diffusion[12]. By contrast, the Indo-Oceanic lineage had the smallest long-term within-lineage THD, consistent with the highly relaxed structure of the MSTree, indicative of genetically diverse strains with few recent transmission events. Between these extreme cases, the Euro-American MSTree was dense,

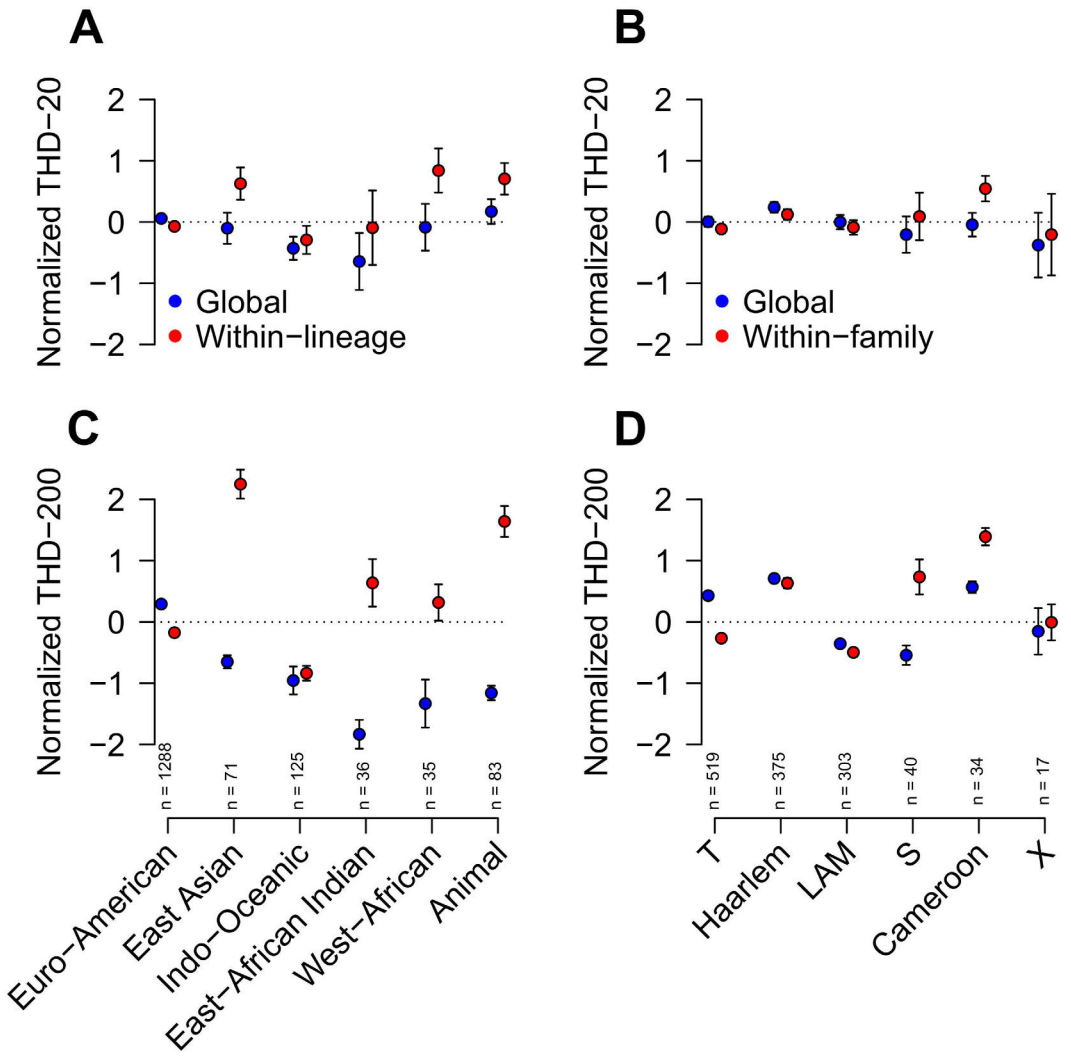


Figure 4. Comparison of time-scaled haplotypic densities (THDs) between MTBC lineages (**A,C**) and spoligotype families within the Euro-American lineage (**B,D**). Short (20y, (**A,B**) and long (200y, (**C,D**) THD timescales were used to reflect short- and long-term evolutionary success, respectively. THDs were computed either with respect to the complete collection of strains (global THD, blue marks) to reflect evolutionary success at the level of the cohort, or independently within each group (within-lineage or -family THDs, red marks) to reflect evolutionary success independent of the global population structure in the collection. Circles denote mean log-THDs; error bars are 95% CI of the mean (not visible when CI is smaller than marker size). Sample size in each group are indicated above the X-axis. Indications of statistical significance were omitted for readability.

with long branches and no obvious central node. Reflecting this diversity, Euro-American strains had the second smallest long-term within-group THDs (Fig. 4C).

Short-term within-group THDs reflected the distribution of closely related haplotypes in each group. These THDs were comparably high in West-African, animal and East-Asian/Beijing lineages, reflecting the large proportion of strains belonging to clusters of identical MIRU-VNTR haplotypes in these groups (Fig. 3). Of note, identical MIRU-VNTR haplotypes in *M. bovis* did not necessarily reflect recent inter-patient transmission events but also contamination by a common source, namely the Bacillus Calmette Guerin vaccine strain[38].

Global THDs, shown as blue markers in Fig. 4, take both the clonality and prevalence of group into account. Expectedly, the highly frequent Euro-American lineage had the highest long-term global THD in spite of its less-than-average within-group THD. Thus, this contrast between global- and within-group THDs highlighted the endemic nature of the Euro-American lineage, both prevalent and diversified, in our setting. Analyses at the sublineage level (Fig. 4D) indicated that the Haarlem and T families mostly contributed to the endemicity of the Euro-American lineage. The Cameroon family, although unfrequent, also had a high THD value, consistent with the previously reported success of this clade in Western Africa[39], from which most patient infected with Cameroon strains originated (n = 17/20, 85%). The East-Asian/Beijing lineage had the second highest long-term global THD in spite of being ranked fourth by decreasing order of prevalence. Interestingly, this lineage also exhibited a high short-term global THD. In line with the radiating MSTree structure observed for this lineage, this

Factor	Short-term THD (20y time-scale)		Long-term THD (200y time-scale)	
	Coeff. (95% CI)[a]	*P*-value	Coeff. (95% CI)	*P*-value
Age at diagnosis (per 10 years)	**−0.04 (−0.06, −0.02)**	**1.8×10^{-4}**	**0.08 (0.06, 0.01)**	**7.8×10^{-16}**
Male sex	0.02 (−0.08, 0.12)	0.73	−0.06 (−0.16, 0.04)	0.21
French-native	0.15 (−0.02, 0.31)	0.08	**0.39 (0.23, 0.55)**	**2.1×10^{-6}**
No. of years in France before diagnosis (non-native patients)	0.00 (−0.01, 0.01)	0.80	0.00 (−0.01, 0.01)	0.92
Collective dwelling	0.07 (−0.21, 0.36)	0.61	0.17 (−0.12, 0.46)	0.26
Occupation[b]	—	0.36	—	**1.6×10^{-3}**
Employed	Reference	—	Reference	—
Retired	−0.21 (−0.56, 0.15)	0.25	**0.40 (0.05, 0.75)**	**2.6×10^{-2}**
Student	−0.29 (−0.73, 0.15)	0.19	**−0.44 (−0.87, −0.01)**	**4.7×10^{-2}**
Unemployed	0.04 (−0.36, 0.45)	0.83	0.06 (−0.34, 0.46)	0.78
Pulmonary infection	**0.17 (0.03, 0.3)**	**2.0×10^{-2}**	**0.24 (0.09, 0.38)**	**1.2×10^{-3}**
AFB-positive sputum	**0.34 (0.18, 0.50)**	**2.0×10^{-5}**	0.15 (−0.01, 0.30)	0.06
Rifampicin resistance	0.14 (−0.22, 0.50)	0.44	−0.13 (−0.49, 0.23)	0.47
Isoniazid resistance	−0.01 (−0.26, 0.25)	0.96	**−0.33 (−0.59, −0.08)**	**9.4×10^{-3}**
Multidrug resistance	0.20 (−0.21, 0.61)	0.33	−0.1 (−0.50, 0.31)	0.64

Table 2. Socio-demographic and disease-related factors associated with short- and long-term time-scaled haplotypic densities (THD) in MTBC-infected patients. [a]Coefficients of linear regression of normalized log-THD, expressed as multiple of standard deviation. Significant coefficients and *P*-values (*t*-test) highlighted in bold. [b]Reported are the model-wise *P*-value of multiple regression model (*F*-test) and category-specific coefficients taking the employed category as reference.

pattern indicates that Beijing strains, although neither prevalent or endemic in our setting, exhibit a high degree of clonality suggestive of a recent epidemic success.

Collectively, these analyses identified the Euro-American strains, mostly in the Haarlem and T families, as being part of the endemic background of tuberculosis in our setting. The results also highlight the recent epidemic success of Beijing strains in spite of their low prevalence.

Factors associated with short- and long-term epidemic success of MTBC strains. Using global THD20 and THD200 as proxies for short- and long-term epidemic success in MTBC strains and their respective lineages, we conducted association studies to identify characteristic features of successful strains and of their infected hosts. Bivariate linear regression analyses detected several such success-associated features (Table 2). Importantly, two of these associations could be considered as positive controls of our analysis. First, smear-positive patients have a well-known higher risk of transmitting disease and of being part of a recent transmission chain, hence their isolates were expected to exhibit higher THD20 values. Second, considering that patients are more likely to harbor strains that are endemic in their region of origin[4,5], isolates from French-native patients were expected to exhibit higher THD200 values. Both associations of THD20 with sputum smear positivity and of THD200 with French-native status had indeed highly significant *P*-values in bivariate analysis. This indicated that THD correctly identified these known and relevant epidemiological processes, in turn suggesting the relevance of this analysis for detecting other associations. Where applicable, we thus examined these associations in more details using stratified analyses and multiple regression models controlling for potential confounders.

Along with sputum smear positivity and pulmonary infection, THD20 correlated with younger age, in line with previous observations that MTBC genotype clustering was more frequent in younger patients[3]. This association was still significant after excluding *M. bovis* strains from the analysis ($P = 6.7 \times 10^{-3}$), indicating a link with patient-to-patient transmission patterns rather than a bias due to Bacillus Calmette Guerin vaccine strain-related infections in infants. However, separate spline regression curves (see Methods) constructed for French-native and non-native patients (Supplementary Fig. S3) indicated that the association pattern of age with THD20 was specific of French-native patients, as THD20 did not change with age in non-native patients. Surprisingly, the association of student status with smaller THD200 retained its amplitude after controlling for age and French-native status, although not significantly so (coefficient −0.49, $P = 0.07$). Among the 201 patients with known occupation and country of origin, 31 (15.4%) were students of which 24 (77.4%) were French-native, suggesting that lower THD200 in MTBC-infected students was not related to a high proportion of non-French-native students in our cohort. Indeed, when restricting the analysis to French-native patients (n = 225) and controlling for age, student status was still associated with a lower THD200 (coefficient −1.06 compared to employed patients, $P = 0.019$). Collectively, these results suggest that students are more likely to be infected with MTBC strains that do not belong to the endemic background.

MTBC strains involved in a pulmonary infection exhibited larger THD200s both in bivariate analysis (Table 2) and after controlling for age and French-native status (coefficient 0.28, $P = 3.1 \times 10^{-3}$). The proportion of pulmonary infections varied depending on the continent of origin of the patients ($P < 10^{-5}$, Fisher's exact test), from 64.0% (n = 174/272) in African-born patients to 74.3% (n = 55/74) and 84.3% (n = 210/249) in Asian- and

European-born patients, respectively (other continents omitted due to small sample sizes), and on the lineage of the infecting strain ($P < 10^{-5}$). This proportion was largest in strains of the East-Asian/Beijing lineage (n = 40/50, 80.0%), followed by the West-African (n = 19/25, 76.0%), Euro-American (n = 522/693, 75.3%), Indo-Oceanic (n = 48/74, 64.9%), East-African Indian (n = 10/20, 50.0%) and animal (n = 15/57, 26.3%) lineages. The association of pulmonary infection with THD200 retained significance after controlling for the continent of origin ($P = 0.002$) but not for the phylogenetic lineage ($P = 0.93$; both models also controlled for age). Hence, the bivariate association of THD200 with pulmonary infection mainly resulted from the association between the two most successful lineages in the long-term, namely the Euro-American and East-Asian/Beijing lineages (Fig. 4C), with high proportions of pulmonary infections compared to other lineages. To determine whether pulmonary infection influenced THD200 at the sub-lineage level, regression models controlled for age and French-native status were constructed for each lineage independently. An independent association was still present between pulmonary infection and THD200 in strains of the East-Asian/Beijing lineage (coefficient 0.76, $P = 2.9 \times 10^{-3}$) but not of other lineages, which suggested that the ability of MTBC Beijing strains to cause pulmonary infection influenced their long-term epidemic success.

Discussion

To our knowledge, the THD framework represents the first approach to allow for in-depth joint analysis of epidemic success over time with pathogen- and host-associated factors in highly structured pathogen populations. By applying this approach on a large cohort of TB patients, we identified factors that contributed to the short- or long-term epidemic success of particular strains in a typical low-prevalence, French setting.

Interestingly, associations of bacterial- and host-related factors with epidemic success/THD measures (Table 2) depended on the timescale considered. As a consistent example, short-term THD, hence short-term epidemic success, was associated with positivity of smear sputum, which is well known to impact on patient contagiousness[40]. Sputum positivity, which reflects disease severity[41,42], is thought to be linked to host-related factors, both behavioral or connected to genetic susceptibility[42,43], but perhaps as well to pathogen-related factors[44]. The general causal relationship between long-term THD and epidemic success and pulmonary forms of TB, representing the infectious form of the disease, is also straightforward. More remarkably, the Euro-American and Beijing lineages exhibited both high long-term THD values and rates of pulmonary TB (see Supplementary Fig. S4). Higher rates of pulmonary infection caused by strains of these lineages has been reported previously by Click *et al.* in the US population[45]. This association was reportedly independent of race/ethnicity, HIV status, age and sex, suggesting that it reflected lineage- rather than patient-specific characteristics. Such interpretation is further supported by the striking similarity between per-lineage proportions of pulmonary disease reported by Click *et al.* in their US patient population and those found in our French cohort: 87.0 vs. 80.0%, 86.2 vs. 75.3%, 77.4 vs. 64.9% and 65.7 and 50.0% for the East Asian/Beijing, Euro-American, Indo-Oceanic and East African-Indian lineages, respectively ($R^2 = 0.99$, $P = 0.006$). Taken collectively, these results bring additional support to the hypothesis of MTBC lineage-specific adaptations impacting on disease[1,44,46–49], including the ability to generate active pulmonary TB as a major driving force of MTBC population dynamics[44,50].

Some limitations of our study prevented us to test several hypotheses. In particular, the proportion of missing data was high for several possibly important factors, such as the time of arrival in France of non-native patients (Table 1), and individual risk factors for tuberculosis such as HIV infection or other immunological impairments were not available for analysis. Although ignoring these factors is unlikely to have biased our conclusions regarding the relationship of pulmonary tuberculosis and long-term epidemic success of MTBC lineages, their inclusion in models involving short-term THD could have helped refining our association analysis.

Compared to current maximum-likelihood or Bayesian methods for investigating pathogen demography, THD is less computationally demanding due to its simplicity, potentially allowing for the analysis of larger strain collections. This computational efficiency is linked to the absence of an explicit phylogenetic reconstruction and to the choice of the efficient but approximate infinite alleles model (IAM) to calibrate the bandwidth (see Methods). These methodological choices have practical consequences regarding the applicability of THD for future studies. First, due to the absence of an evolutionary model, THD can handle any type of qualitative genetic data that bears phylogenetic information, such as minisatellites or DNA polymorphisms. Second, although the IAM model is reasonably accurate for recent TMRCAs[51], it does not consider locus homoplasy and tends to underestimate TMRCA when genetic distance increases, as illustrated in Supplementary Fig. S5. Hence, THD analyses should be restricted to relatively recent timescales so that locus homoplasy can be safely ignored. We empirically suggest that the chosen timescale should not yield a median genetic distance greater than one-third of the number of loci (corresponding to a maximal timescale of ≈400y for 15-loci MIRU-VNTR; Supplementary Fig. S5). Finally, one should keep in mind that typing methods of routine use such as 15-loci MIRU-VNTR convey much less information than, e.g., whole genome sequences, and that a large sample size (say, $n > 100$) is desirable to compensate for the uncertainty inherent to low-resolution data (Fig. 2 and Supplementary Fig. S5).

In summary, our results describe how the interplay of MTBC lineage specificities and host risk factors contribute to the large-scale population dynamics of MTBC in a low-prevalence setting. Analyses focused on longer or shorter timescales confirmed the potential driving forces of the epidemic success of MTBC such as the propensity to cause transmissible, pulmonary disease in the long run and sputum-positive infections in the short run. Such approach could be used more generally to infer the epidemic success of pathogens with widely available typing data, including SNPs, and to reveal relevant associations with factors suspected to influence the course of an epidemic over time.

Methods

Timescaled haplotypic density. We consider the problem of using kernel density estimation to assign a measure of density to a haplotype, represented as a vector of markers, relative to a set of other haplotypes. After providing the required definitions, we briefly expose the kernel function, the computation of the bandwith based on a timescale parameter, and we provide a synthetic overview of THD computation. Source code of the software implementation of these methods for the R platform is available in the Supplementary Note.

Let X be a sample of n haplotypes defined over m markers, represented as an $(n \times m)$ data matrix, and let y be a haplotype of interest not in X. For each haplotype x_i in X, let h_i be the genetic distance from y to x_i, i.e. the number of differences between x_i and y. A genetic distance h is associated with a kernel density (formally, a probability) $k(h|b,m)$ under the truncated geometric distribution with bandwidth b (formally, the failure probability of a Bernoulli trial) and truncation limit m. This distribution has probability mass function $k(h|b,\ m) = \left(\frac{1-b}{1-b^{m+1}}\right)b^h$. The bandwidth b is a real number between 0 and 1. The density associated with a given distance h is proportional to b^h, which illustrates how the bandwidth controls the influence of h on the density: for each additional difference between y and x, the density is multiplied by b. Reducing b, thus, accelerates the decrease of the density for larger numbers of differences. Finally, the haplotypic density $K(y|X,b,m)$ of y with respect to X is the average of the n densities associated with the distances from y to each x_i in X,

$$K(y|X,\ b,\ m) = \frac{1}{n}\sum_{i}^{n} k(h_i|b,\ m) = \frac{1}{n}\left(\frac{1-b}{1-b^{m+1}}\right)\sum_{i}^{n} b^{h_i}$$

Because b is a dimensionless constant, its choice is not intuitive. To circumvent this issue, we exploit the existence of a one-to-one relationship between genetic distance h and the maximum-likelihood estimate of the TMRCA t under the infinite alleles model (IAM)[51], which assumes that the m haplotype markers lie on a non-recombining DNA segment, that they evolve independently with a common evolutionary rate μ, and that at most one change per marker occurred in both lineages since their MRCA. Assuming that μ is known, the IAM model allows to replace the bandwidth with a more intuitive timescale parameter t_{50}, or tMRCA50, which is the TMRCA such that haplotypes with shorter TMRCAs account for 50% of the density. Practically, we solve the IAM model relation $t = \log[m/(m-h)]/2\mu$ for h to obtain $h = (1-e^{-2\mu t})m$. This relation allows to associate a (possibly non-integer-valued) distance h_{50} with the chosen timescale t_{50}. From the definition of t_{50}, it follows that h_{50} is the median of a truncated geometric distribution whose bandwith b_* must be determined. From the cumulative probability function of the truncated geometric distribution with parameters b and m, $P(H \leq h|b,\ m) = \frac{1-b^h}{1-b^m}$, it follows that if h_{50} is the median of the continuous form of the distribution with bandwidth b_* then b_* must satisfy $\frac{1-b_*^{h_{50}}}{1-b_*^{m}} = \frac{1}{2}$, an equation which we can solve for b_* numerically (as no closed-form solution exists) using a root-finding algorithm over the [0,1] interval.

THD computation steps can be summarized as follows: (i) determine parameters m (number of markers), μ (evolutionary rate) and t_{50} (timescale); (ii) associate the timescale with a median distance h_{50}; (iii) determine the corresponding bandwidth b_*; and iv) for each haplotype of interest, compute THD as the average kernel density under the truncated geometric distribution with bandwidth b_* and truncation limit m.

Summarizing, scaling and normalizing THD. Because THDs are probabilities, aggregate statistics for groups of isolates should use products to represent the joint likelihood of isolates in the group. As a consequence, we use geometric means rather than arithmetic means, and transform THDs to logarithms before inclusion in linear models. Because THD estimates are inversely proportional to the number of loci m and sentitive to the timescale we propose two modifications to ease comparability. First, THDs can be multiplied by the number of loci. These scaled THDs are invariant relative to m (Fig. 2 and Supplementary Fig. S1), which might facilitate comparison between THDs with similar timescale but obtained with different methods, e.g. minisatellite typing vs. DNA sequencing. Second, log-THDs can be centered and scaled relative to the population under study. These normalized THDs are multiples of standard deviations from the mean of the population, similar to Z-scores. They are not comparable across studies, as they depend on a given population, but they can be compared between different timescales.

Simulation experiments. Simulation of minisatellite-based haplotypes, evolving under the stepwise mutation model using a continuous-time sequential Markov coalescent approximation, were conducted by means of Fastsimcoal2 software[22]. Scenario parameters were set as indicated in text. Simulated haplotypes were imported into the R software environment for THD computation and further analyses.

Ethics statement. This retrospective, cross-sectional, observational multicentric study was approved by the Comité de Protection des Personnes Sud-Est IV under no. DC-2011-1306. Written consent of participants was not obtained, in accordance with French regulations, due to anymous treatment of data and the non-interventional nature of the study.

Patient population and collection of data. Patients were identified retrospectively from the surveillance database of the ORAM, a regional collaborative surveillance system active since 2005 whose participants include: i) the microbiology laboratories of the three university hospitals of the Rhône-Alpes region, namely Lyon, Grenoble and Saint-Etienne, as well as other microbiology laboratories in charge of tuberculosis diagnosis; ii) the Agence Régionale de Santé (ARS) to which all TB diagnoses are notified by practitioneers as part of the French programme for tuberculosis surveillance; and iii) the Centre de Lutte Anti-Tuberculeuse (CLAT) which

is in charge of the identification of contact cases and of the long-term follow-up of tuberculosis patients after hospital discharge. MTBC strains isolated by the participating laboratories are routinely referred to reference laboratories for molecular typing, including spoligotyping since 2005 and 15-loci MIRU-VNTR typing since 2008. Typing methods were consistent in all ORAM laboratories. Spoligotyping was performed as described elsewhere[29]. Spoligotypes were compared to the SpolDB4 database of Institut Pasteur and assigned to lineages and sublineages[35].

Patients were eligible if the tuberculosis diagnosis was notified to the ARS between 2008 and 2014 and if their infecting strain had been isolated and typed (n = 1,746). Patients whose MTBC strain had ambiguous MIRU-VNTR profile (i.e., undefined number of repeats at any of the 15 loci; n = 105) were excluded. Demographic data extracted from the database included gender, age at the time of diagnosis, year of isolation of the MTBC strain, country of birth, occupation (employed, unemployed, student or retired), and collective dwelling (including nursing home, group home, prison and refugee camp). Disease-related data included disease location, sputum smear positivity and phenotypic rifampin and isoniazide resistance. Disease was classified as pulmonary or exclusively extra-pulmonary. Patients with exclusively extra-pulmonary disease were those with at least 3 sputum samples with negative MTBC culture result. If <3 sputum samples were taken, disease location was considered unknown (Table 1).

Population structure analysis. MSTrees were computed based on the 15-loci MIRU-VNTR haplotypes using BioNumerics 7.5 (Applied Maths, St Martens-Latem, Belgium). An MSTree is a connected undirected graph selected to minimize the sum of marker differences over all links between haplotypes, enabling the graphical representation of quantitative relationships between MIRU-VNTR haplotypes. Independent MSTrees, one per major lineage, were constructed.

Statistical analysis. Association studies of socio-demographic, disease-related and microbiological parameters with THD measures were conducted by means of multiple linear regression models with log-THD as the response variable. Control for confounding was achieved by including potential confounders, indicated in text as appropriate, as covariates. Acceptability of linear regression assumptions was assessed by visual inspection of residual distributions and quantile-quantile plots. In line with the exploratory nature of the study, no *P*-value correction for multiple testing was applied. The significance threshold was set at 0.05 for all tests. Spline regression curves based on cubic spline interpolation with automatic selection of the smoothing parameter were used to visualize possible non-linear relationships between variables. All computations were performed using R software version 3.0.1 Good Sport (The R Foundation for Statistical Computing, Vienna, Austria).

References

1. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45,** 1176–1182 (2013).
2. The Lancet Respiratory Medicine. Changing minds about tuberculosis. *Lancet Respir Med* **3,** 901 (2015).
3. Allix-Béguec, C., Supply, P., Wanlin, M., Bifani, P. & Fauville-Dufaux, M. Standardised PCR-based molecular epidemiology of tuberculosis. *Eur. Respir. J.* **31,** 1077–1084 (2008).
4. Reed, M. B. *et al.* Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J. Clin. Microbiol.* **47,** 1119–1128 (2009).
5. Fallico, L. *et al.* Four year longitudinal study of *Mycobacterium tuberculosis* complex isolates in a region of North-Eastern Italy. *Infect. Genet. Evol.* **26,** 58–64 (2014).
6. Walker, T. M. *et al.* Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* **2,** 285–292 (2014).
7. Caminero, J. A. *et al.* Epidemiological evidence of the spread of a *Mycobacterium tuberculosis* strain of the Beijing genotype on Gran Canaria Island. *Am. J. Respir. Crit. Care Med.* **164,** 1165–1170 (2001).
8. Pena, M. J. *et al.* Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches. *Thorax* **58,** 618–622 (2003).
9. Pybus, O. G., Rambaut, A. & Harvey, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155,** 1429–1437 (2000).
10. Strimmer, K. & Pybus, O. G. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* **18,** 2298–2305 (2001).
11. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).
12. Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47,** 242–249 (2015).
13. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4,** e1000160 (2008).
14. Lee, R. S. *et al.* Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* **112,** 13609–13614 (2015).
15. Luo, T. *et al.* Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proc Natl Acad Sci USA* **112,** 8136–8141 (2015).
16. Jamrozy, D. & Kallonen, T. Genome watch: Looking at Beijing's skyline. *Nat Rev Micro* **13,** 528–528 (2015).
17. Ho, S. Y. W. & Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour* **11,** 423–434 (2011).
18. Heller, R., Chikhi, L. & Siegismund, H. R. The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS One* **8,** (2013).
19. Cantinelli, T. *et al.* 'Epidemic Clones' of *Listeria monocytogenes* Are Widespread and Ancient Clonal Groups. *J Clin Microbiol* **51,** 3770–3779 (2013).
20. Biek, R., Pybus, O. G., Lloyd-Smith, J. O. & Didelot, X. Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* (*Amst.*) **30,** 306–313 (2015).
21. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* **33,** 1065–1076 (1962).
22. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27,** 1332–1334 (2011).
23. Reyes, J. F. & Tanaka, M. M. Mutation rates of spoligotypes and variable numbers of tandem repeat loci in *Mycobacterium tuberculosis*. *Infect. Genet. Evol.* **10,** 1046–1051 (2010).

24. Supply, P., Niemann, S. & Wirth, T. On the mutation rates of spoligotypes and variable numbers of tandem repeat loci of *Mycobacterium tuberculosis*. *Infect. Genet. Evol.* **11,** 251–252 (2011).
25. Ragheb, M. N. *et al.* The mutation rate of mycobacterial repetitive unit loci in strains of *M. tuberculosis* from cynomolgus macaque infection. *BMC Genomics* **14,** 145 (2013).
26. Allix-Béguec, C. *et al.* Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. *J. Clin. Microbiol.* **52,** 164–172 (2014).
27. Pepperell, C. S. *et al.* The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog.* **9,** e1003543 (2013).
28. Pichat, C. *et al.* Combined Genotypic, Phylogenetic, and Epidemiologic Analyses of *Mycobacterium tuberculosis* Genetic Diversity in the Rhône Alpes Region, France. *PLoS ONE* **11,** e0153580 (2016).
29. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35,** 907–914 (1997).
30. Supply, P. *et al.* Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **44,** 4498–4510 (2006).
31. Comas, I., Homolka, S., Niemann, S. & Gagneux, S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE* **4,** e7815 (2009).
32. Mokrousov, I. *et al.* Analysis of the allelic diversity of the mycobacterial interspersed repetitive units in *Mycobacterium tuberculosis* strains of the Beijing family: practical implications and evolutionary considerations. *J. Clin. Microbiol.* **42,** 2438–2444 (2004).
33. Mokrousov, I. *et al.* Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. *Genome Res.* **15,** 1357–1364 (2005).
34. Mokrousov, I. Insights into the origin, emergence, and current spread of a successful Russian clone of *Mycobacterium tuberculosis*. *Clin. Microbiol. Rev.* **26,** 342–360 (2013).
35. Brudey, K. *et al. Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* **6,** 23 (2006).
36. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103,** 2869–2873 (2006).
37. Shabbeer, A. *et al.* TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect. Genet. Evol.* **12,** 789–797 (2012).
38. Cuello-García, C. A., Pérez-Gaxiola, G. & Jiménez Gutiérrez, C. Treating BCG-induced disease in children. *Cochrane Database Syst Rev* **1,** CD008300 (2013).
39. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48,** 1535–1543 (2016).
40. Musher, D. M. How contagious are common respiratory tract infections? *N. Engl. J. Med.* **348,** 1256–1266 (2003).
41. Pagaoa, M. A. *et al.* Risk factors for transmission of tuberculosis among United States-born African Americans and Whites. *Int. J. Tuberc. Lung Dis.* **19,** 1485–1492 (2015).
42. Nebenzahl-Guimaraes, H., Verhagen, L. M., Borgdorff, M. W. & van Soolingen, D. Transmission and Progression to Disease of *Mycobacterium tuberculosis* Phylogenetic Lineages in The Netherlands. *J. Clin. Microbiol.* **53,** 3264–3271 (2015).
43. Casanova, J.-L. & Abel, L. Genetic Dissection of Immunity to Mycobacteria: The Human Model. *Annual Review of Immunology* **20,** 581–620 (2002).
44. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin. Immunol.* **26,** 431–444 (2014).
45. Click, E. S., Moonan, P. K., Winston, C. A., Cowan, L. S. & Oeltmann, J. E. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clin. Infect. Dis.* **54,** 211–219 (2012).
46. Parwati, I., van Crevel, R. & van Soolingen, D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* **10,** 103–111 (2010).
47. Sarkar, R., Lenders, L., Wilkinson, K. A., Wilkinson, R. J. & Nicol, M. P. Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. *PLoS ONE* **7,** e43170 (2012).
48. Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA* **99,** 3684–3689 (2002).
49. Gonzalo-Asensio, J. *et al.* Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc. Natl. Acad. Sci. USA* **111,** 11491–11496 (2014).
50. Dale, J. W. *et al.* Origins and properties of *Mycobacterium tuberculosis* isolates in London. *J. Med. Microbiol.* **54,** 575–582 (2005).
51. Walsh, B. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* **158,** 897–912 (2001).

Acknowledgements

We are indebted to all the microbiologists, physicians and technical staff who contributed to the ORAM, transmitted information and referred MTBC strains to reference laboratories for typing. We also thank E. Rivollier (ULAT of Saint-Etienne, France), T. Ferry, T. Perpoint and F. Valour (Lyon University Hospital) for help with data collection, and S. Mona, S. Boitard, M. Veuille (Muséum National d'Histoire Naturelle, France) and S. Niemann (Research Center Borstel, Germany) for fruitful discussion. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

J.P.R. and T.W. designed the study; O.D., C.P., G.C., A.S.R.B., G.B., C.G.B., S.B., A.C., V.J., I.F., M.P.M. and F.A. collected data; J.P.R., M.B. and C.P. performed the experiments; J.P.F. and G.L. contributed reagents/tools; J.P.R. and T.W. analyzed the data; J.P.R., M.B., F.A., P.S. and T.W. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: Veronique Jacomo is employed by Biomnis Laboratories, a commercial company. The authors declare no competing financial interests regarding the present work.

How to cite this article: Rasigade, J.-P. *et al.* Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis. *Sci. Rep.* **7**, 45326; doi: 10.1038/srep45326 (2017).

Chapitre 6.

Les fluctuations des patrons de migrations humaines ont forgé la structure de population globale de *Mycobacterium tuberculosis* en France

Changing patterns of human migrations shaped the global population structure of *Mycobacterium tuberculosis* in France

Maxime Barbier,[1,2] Oana Dumitrescu,[3,4] Catherine Pichat,[4] Gérard Carret,[4] Anne-Sophie Ronnaux-Baron,[5] Ghislaine Blasquez,[5] Christine Godin-Benhaim,[6] Sandrine Boisset,[7,8] Anne Carricajo,[9] Véronique Jacomo,[10] Isabelle Fredenucci,[4] Michèle Pérouse de Montclos,[4] Charlotte Genestet,[3] Jean-Pierre Flandrois,[4,11] Florence Ader,[3,12] Philip Supply,[13] Gérard Lina,[3,4] Thierry Wirth[1,2] & Jean-Philippe Rasigade[1,2,3,4*]

[1]Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Université Pierre et Marie Curie, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France. [2]Laboratoire Biologie Intégrative des Populations, Ecole Pratique des Hautes Etudes, PSL Research University, Paris, France. [3]Centre International de Recherche en Infectiologie, CIRI, University of Lyon, France. [4]Institut des Agents Infectieux, Hospices Civils de Lyon, Lyon, France. [5]Comité Départemental d'Hygiène Sociale, CLAT69, Lyon, France. [6]Agence Régionale de Santé Auvergne-Rhône-Alpes, Lyon, France. [7]Laboratoire de Bactériologie, Institut de Biologie et de Pathologie, CHU de Grenoble, Grenoble, France. [8]Laboratoire TIMC-IMAG, UMR 5525 CNRS-UJF, UFR de Médecine, Université Grenoble Alpes, Grenoble, France. [9]Laboratoire des Agents Infectieux et d'Hygiène, CHU de Saint-Etienne, Saint-Etienne, France. [10]Laboratoire Biomnis, Lyon, France. [11]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, University of Lyon, France. [12]Service des Maladies Infectieuses et Tropicales, Hôpital de la Croix-Rousse, Hospices Civils de Lyon, Lyon, France. [13]INSERM U1019, CNRS-UMR 8204, Center for Infection and Immunity of Lille, Institut Pasteur de Lille, Université de Lille, Lille, France

*Address correspondence to Jean-Philippe Rasigade, jean-philippe.rasigade@univ-lyon1.fr

Counts: Abstract, 150 words; Text excluding Methods: 2,620 words; Methods: 710 words;

References: 30

Abstract

Mycobacterium tuberculosis complex (MTBC) exhibits a structured phylogeographic distribution worldwide linked with human migrations. We sought to infer how the interactions between distinct human populations shape the global population structure of MTBC on a regional scale. We applied the recently described timescaled haplotypic density (THD) technique on 638 minisatellite-based MTBC genotypes from French tuberculosis patients. THD with a long-term (200y) timescale indicated that MTBC population in France had been mostly influenced by interactions with Eastern and Southern Europe and, to a lesser extent, Northern and Middle Africa, consistent with historical migrations favored by geographic proximity or commercial exchanges with former French colonies. Restricting the timescale to 20y, THD identified a sustained influence of Northern Africa, but not Europe where tuberculosis incidence decreased sharply. Evolving interactions between human populations, thus, measurably influence the local population structure of MTBC. Relevant information on such interactions can be inferred using THD from MTBC genotypes.

Tuberculosis, one of the oldest diseases known to humanity, is an ongoing public health threat in many low-income countries and a re-emerging disease in several higher-income countries[1–3]. The causative agent of tuberculosis, *Mycobacterium tuberculosis* complex (MTBC), is thought to have co-evolved with modern humans since their expansion out of Africa ~60,000y ago[4,5], a situation also encountered in other bacterial pathogens such as *Helicobacter pylori*[6–8]. The dispersal of human populations was accompanied with a genetic diversification of MTBC into distinct lineages whose current distribution exhibits variable degrees of geographic specificity[9,10]. Based on this specificity, the assignment of novel MTBC strains to lineages using molecular methods reveals the possible geographic origins of the strain's ancestry[11]. Such approaches are essential since the connection between MTBC lineages and patients origin has a practical importance for MTBC studies in low-incidence countries where a large proportion of tuberculosis cases are imported from abroad[12,13].

More recently, phylogeographic studies of MTBC involving population genetics analysis methods have stressed the impact of social phenomena such as migrations on the evolution and population structure of MTBC lineages[10,14]. Most previous studies of MTBC phylogeography have examined continental-scale epidemics. Each country or region has a unique history of ancient and recent interactions with foreign populations that might have shaped the local MTBC population structure, but such specific events might go undetected in large-scale studies. To facilitate more local, fine-grained phylogeographic studies without the cost and time constraints of international sampling, one might consider MTBC strains infecting foreign-born patients as proxies of the circulating strains in the patients country of origin[11,15]. Under this assumption, a

representative sample of MTBC isolates from an area of interest should enable to infer past transmission events involving the countries of origin of non-native patients.

Based on this rationale, we performed a comprehensive phylogeographic analysis of global MTBC strains circulating in the Rhônes-Alpes region of France, a low-incidence country where two thirds of cases involve foreign-born patients[13]. Our primary objective was to contribute an in-depth description of the influence of past interactions of French and foreign populations on the current population structure of MTBC in our region. Our secondary objective was to empirically assess the relevance of using MTBC strains from migrant patients to infer phylogeographic information on a local scale. We investigated a cohort of 638 tuberculosis patients whose infecting strains were genotyped using spoligotyping[16] and variable number of tandem repeats (VNTR)[17] methods. Genotype-based assignment of isolates to lineages was analyzed in combination with the country of origin of patients. To detect temporal changes in transmission patterns, we built upon a recently published phylogenetic analysis method, namely timescaled haplotypic density (THD)[18], to estimate the intensity of MTBC exchanges between native and foreign populations over short- and long-term time scales and we interpreted the inferred results in relation with known historical and sociological events.

Results

Study population. Patients were retrieved from the database of the Observatoire Rhône-Alpin des Mycobactéries (ORAM), a regional network of healthcare institutions involved in the monitoring of tuberculosis in the French Rhône-Alpes region, an area with ~6.5 million inhabitants and a population density of 150 inhabitants per km². Patients were eligible for inclusion if: (1) tuberculosis was diagnosed from 2008 to 2014; (2) their country of birth was known; and (3) their infecting strain was available for genotyping using spoligotyping and mycobacterial interspersed repetitive unit (MIRU) VNTR typing with 15 loci. Patients from America and Oceania were excluded due to small sample sizes, as well as patients infected with *M. bovis*, *pinnipedii* and *microti* due to the specific transmission routes of these species (see Methods). Six hundreds and thirty-eight patients were included in the final analysis. The median age was 42y (inter-quartile range 29-63y), the m/f sex ratio was 1.59 and 206 patients (32.3%) were born in France (see Supplementary Table S1 for detailed information). The distribution of patient geographic origins is indicated in Fig 1.

Phylogeographical structure of MTBC in the French population. Spoligotype-based MTBC lineages and families, inferred from the SPOLDB4 database, exhibited a highly structured phylogeographic distribution (Fig. 1). Within the Euro-American lineage, the T, Haarlem and LAM families were strongly associated with a native or North-African origin. Indo-Oceanic and East-African Indian lineages were associated with Asian or East-African origin, while the East-Asian / Beijing lineage was mostly present in Asian and West-Asian patients and, to a lesser extent, in East-European patients, consistent with previous reports of the distribution of these lineages [11,14]. Other lineages and families had more specific distributions, such as the Cameroon

family found in Western- and Middle-African-born patients, and the Western-African lineage associated with a Western-African origin. Collectively, these results were in line with previous reports that migrant patients tend to be preferentially infected with endemic MTBC strains of their country of origin[11,15].

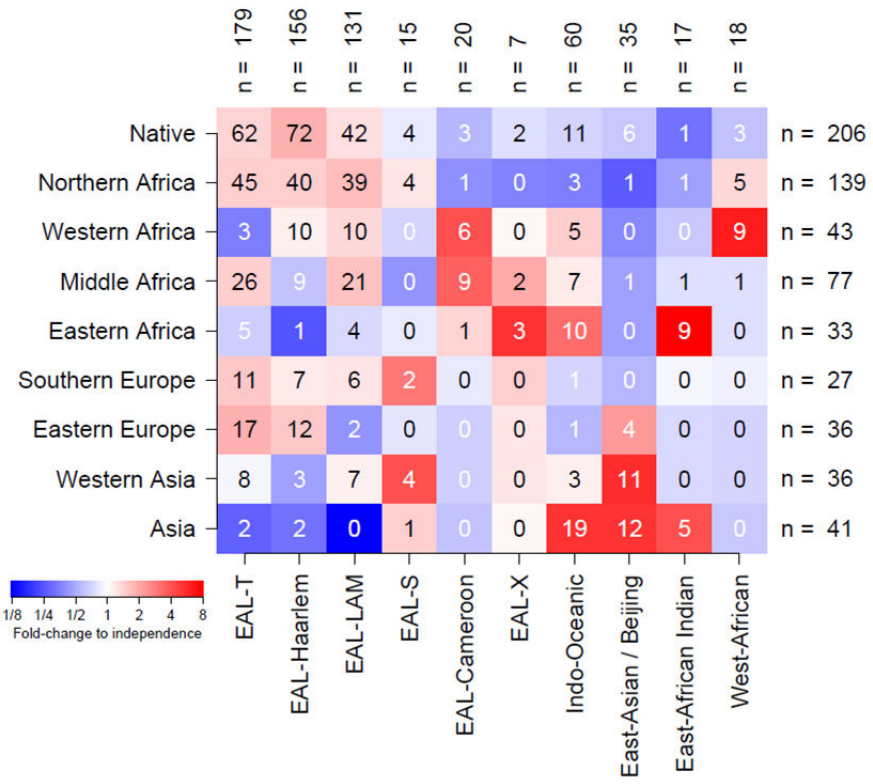


Figure 1. Association heatmap of major MTBC lineages and spoligotype families with patient's region of origin. Shown are the no. of samples in each category, with row- and column-wise sample sizes indicated above and on the right of the heatmap. Colors indicate strength and direction (from blue, strongly negative, to red, strongly positive) of the association between lineage / spoligotype family and region of origin, expressed as fold-change of the observed count in each category relative to the expected count under the hypothesis of independence. Laplace smoothing was applied to proportions to avoid zero fold-changes for zero counts. Spoligotype families belonging to the Euro-American lineage are prefixed with EAL. Other lineages are designated by lineage name.

Timescaled haplotypic density as a measure of pathogen exchange between populations. We used THD based on MIRU haplotypes[17,19] to quantify the intensity of MTBC exchanges between patient populations, taking migrant patients as representative of the population of their country of origin. THD is a recently described population genetics technique that analyzes the population structure of a pathogen sample to assign estimates of transmission success to individual isolates, or haplotypes[18]. THD quantifies the genetic proximity of an isolate with respect to a population with a tuning parameter, the timescale, that restricts the analysis to a given period of time before present, progressively ignoring events older than the timescale. The THD measure was shown in simulations and in an empirical study of MTBC to reflect the number of common ancestors of the isolate and the sampled population in the timescale, which in turn reflects the density of transmission events in the ancestry of the isolate, that is, the isolate's transmission success.

In its original application[18], THD assigns a measure of success to an isolate based on its similarity to other isolates in the whole dataset. If isolates are clustered into groups and THD similarity is measured from isolates in a group relative to those in another group (rather than to the whole dataset), then this similarity reflects the intensity of isolate exchanges between the groups over the considered timescale. Based on this rationale, we generalize the use of THD to groups of isolates and call the resulting measure cross-THD. Here, isolates are grouped according to patient's geographic origin, allowing to perform fine-grained phylogeographic analyses focused on specific timescales. To ease comparisons between groups, cross-THDs are normalized so that they add to unity (see Methods). The cross-THD score of a group relative to itself, termed self-THD, reflects the intensity of exchanges within the group relative to

136 exchanges with other groups. Self-THD, thus, reflects the group's isolation relative to other
137 groups.

138 We examined exchanges of MTBC strains between populations in our cohort using 20y
139 (short-term) and 200y (long-term) THD timescales, consistent with our previous work on MTBC
140 MIRU haplotypes[18]. For each timescale, cross- and self-THDs were computed across all regions
141 of origin including France (native patients). This analysis revealed several important
142 phylogeographic characteristics in our cohort (Fig. 2). Self-THD scores (identified with asterisks
143 in the figure) over the 20y timescale were lowest in patients born in France, Southern Europe
144 and Eastern Europe (21.6%, 22.0% and 22.4%, respectively), suggesting that MTBC population
145 structures in these regions have been the most strongly influenced by exchanges with other
146 regions. Opposite to this situation, Eastern Africa and Asia had higher self-THD (34.5% and
147 30.1%, respectively) indicating more limited pathogen exchanges with the other regions
148 represented in our cohort. This pattern was even more pronounced over the 200y timescale,
149 with self-THDs going down to ~15% for France and Southern Europe, but going up to more than
150 40% for Asia.

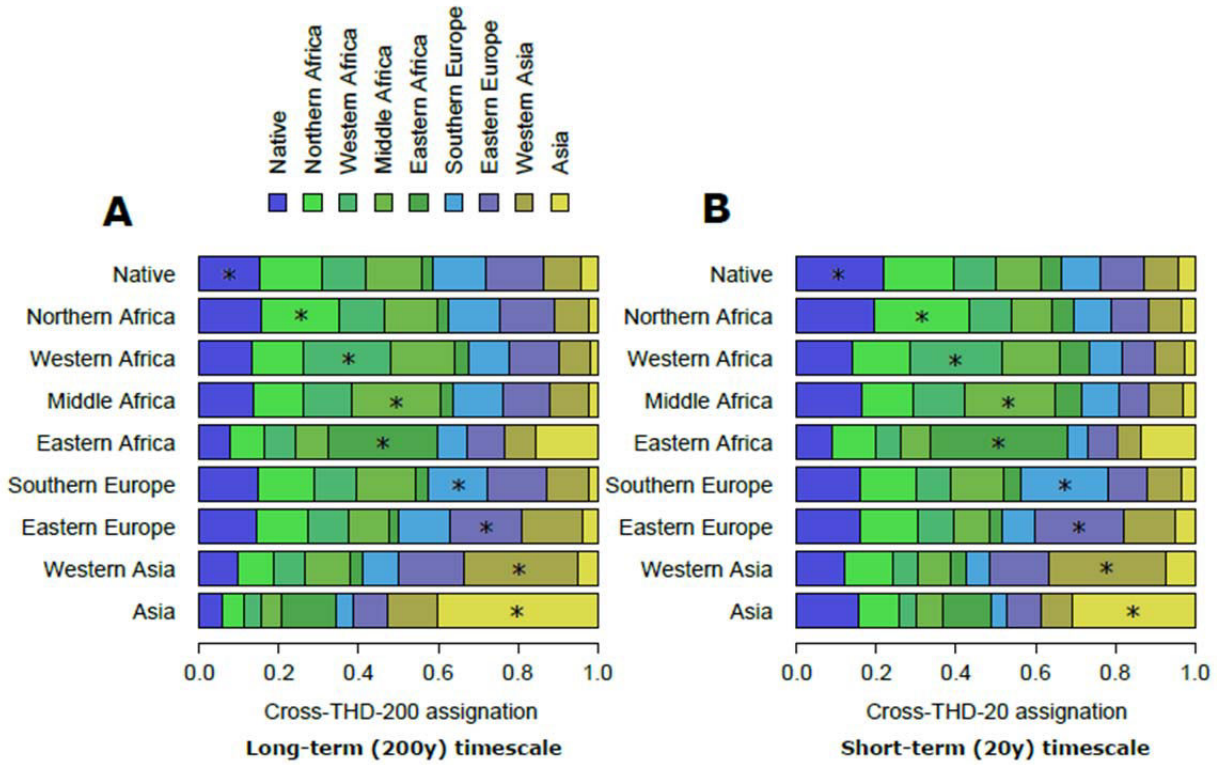


Figure 2. Phylogeographic structure of MTBC strains from the Rhône-Alpes region of France. Cross-THDs with timescales of 200y (A) and 20y (B) were computed based on 15-loci MIRU-VNTR haplotypes in 638 isolates. In each row, cross-THDs of all geographical groups are shown relative to the reference group indicated on the left. Cross-THDs estimate the relative intensity of MTBC exchanges between the reference and the other groups over the indicated timescale. An asterisk indicates self-THD, which estimates the relative intensity of exchanges within the reference group of the row.

Patterns of MTBC exchanges with the French population changed over time. To visualize relationships between the population structures of all geographic regions, cross-THDs were transformed into dissimilarity measures (see Methods) and regions were projected as points on a plane by means of metric multidimensional scaling (MDS; Fig. 3). Using both the 20y and 200y timescales, the French-native group was projected onto the most central position on the first MDS plane, which suggested that this group exhibited on average the largest similarity with all other groups, consistent with its low self-THD score. MDS projection highlighted strong similarities of MTBC population structure between France, Eastern and Southern Europe over the 200y timescale, with a gradient of similarity consistent with the geographical disposition of the contributing regions, from Western and Middle Africa to Northern Africa, Southern Europe and France, while Eastern Africa, Western Asia and Asia were projected further away from the French-native group. After focusing the analysis to the more recent past by using a 20y timescale, several changes occurred in the MDS projection: although the gradient between non-Eastern African regions remained, Northern Africa became the closest region to France, while Southern and Eastern Europe were projected away from France, suggesting changes in the intensity of MTBC exchanges between these regions compared to the long-term timescale.

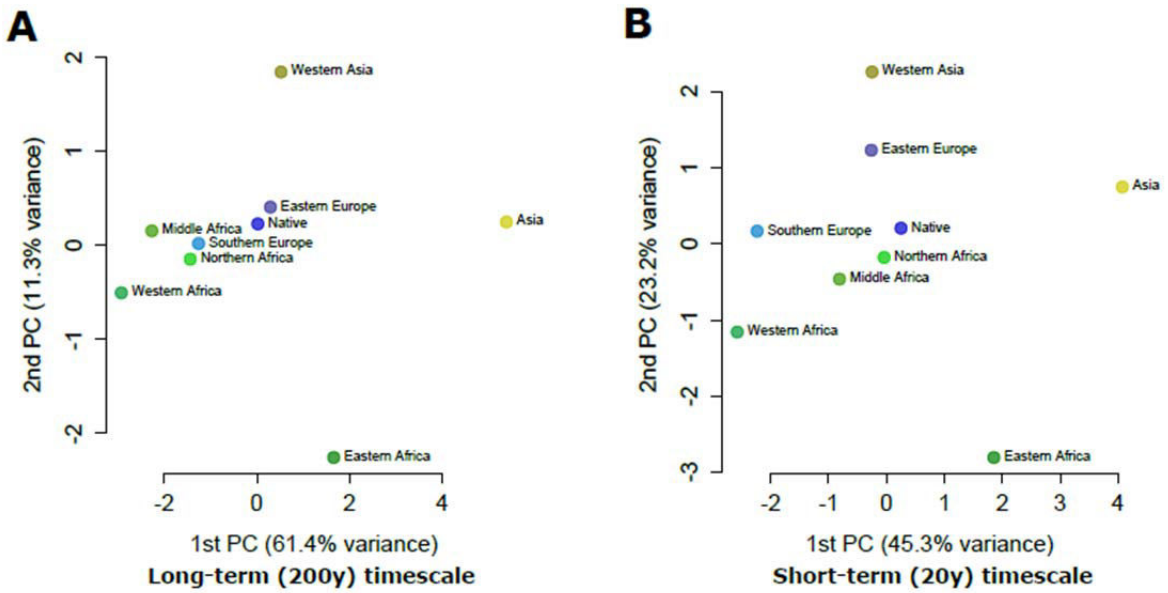


Figure 3. THD-based similarity between MTBC population structures using long-term (A) and short-term (B) timescales. Phylogenetic proximities between regions of origin were estimated by transforming pairwise cross-THDs into distances visualized using multidimensional scaling. PC, principal component.

182 The MDS projections took into account all pairwise similarities between regions without
183 focusing only on similarities with the French-native group. To specifically examine the
184 contribution of each region to the population structure of strains infecting French-native
185 patients, ignoring other pairwise similarities, cross-THDs of each region relative to the French
186 native group were expressed as fold-changes to the self-THD of the native group and displayed
187 on a map (Fig. 4D and F). This analysis confirmed the major contribution of exchanges with
188 European and non-Eastern African populations to the long-term population structure of MTBC
189 in French-native patients, as well as the relative decline of the contribution of most regions
190 excepted North Africa when narrowing THD focus to the short-term 20y timescale.

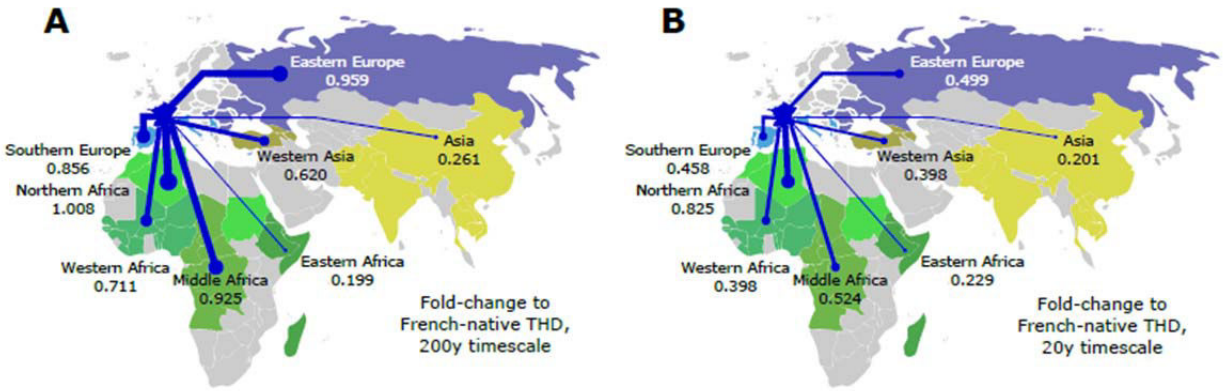


Figure 4. Long- and short-term contributions of interactions between France and other regions to the local MTBC population structure. Line widths are proportional to fold-change of cross-THDs relative to self-THD in French-native patients, using timescales of 200y and 20y (A and B, respectively). Maps were prepared with Inkscape v0.91 software (https://inkscape.org) and adapted from public domain vectorized map file accessible at: https://commons.wikimedia.org/wiki/File:BlankMap-World6.svg.

Discussion

In this analysis of 638 tuberculosis patients from a low-incidence area, we identified distinct patterns of international MTBC transmission that evolved with time and highlighted how the history of a local population contributes to shape its MTBC population structure. Our results also demonstrated how the THD technique, which was developed initially to estimate the transmission success of single isolates[18], can be applied to phylogeographic analyses in a straightforward fashion.

Comparisons of long- and short-term cross-THDs in MTBC strains of French-native and non-native patients unraveled patterns of genetic similarities suggesting how past and recent contacts of the French population with regions of varying MTBC prevalences have influenced the current MTBC population structure. Regions that contributed the most to this structure over a 200y timeframe were either geographically close to France, such as Southern and Eastern Europe, or included former French colonies, such as Northern, Middle and Western Africa (Fig. 3). These inferences are consistent with historical records, documenting that tuberculosis transmission between inhabitants of metropolitan France and African territories increased rapidly in the first half of the 20th century, favored by economic exchanges and the transportation of military recruits, especially during the course of the First World War[20].

Shortening the timeframe to 20y, the similarities in these European and African regions with French MTBC were nearly halved (e.g., from 0.96 to 0.50 and from 0.86 to to 0.458 for Eastern and Southern Europe, respectively; Fig. 4) with the notable exception of the Northern African region. These patterns correlate with known changes in both the intensity of contacts

220 with metropolitan France and in the regional prevalence of MTBC. In Middle and Western
221 Africa, MTBC prevalence has remained high[3], however the intensity of contacts with the French-
222 native population has decreased sharply after the decolonization movement that followed the
223 Second World War. An opposite situation could prevail for Southern and Eastern Europe; these
224 regions remain the major source of immigration to France[21], however they benefited from a
225 sharp decrease in MTBC prevalence in the long term which has probably contributed to lower
226 the transmission. Finally, contacts of French inhabitants with Northern African countries such as
227 Algeria, Morocco and Tunisia have remained very frequent[21,22] and MTBC prevalence in these
228 countries is still high[23], which might explain why the cross-THD contribution of Northern Africa
229 did not decrease substantially from the 200y to the 20y timeframe.

230 Importantly, our analyses relied on the assumption that migrant patients could be taken
231 as representatives of the population of their country of origin, with respect to MTBC molecular
232 epidemiology. This assumption is violated when migrant patients become infected during travel
233 or once in France or when native patients become infected when abroad. Patients born outside
234 Europe and Northern Africa exhibited an MTBC population structure that was clearly distinct
235 from that of native patients (Fig 1). This indicated that infection in France was rare in this
236 migrant population, as was expected given the lower tuberculosis prevalence compared to Asia
237 and Africa. The proximity, however, between population structures found in patients born in
238 France and Northern Africa might have been enhanced by two sociological peculiarities. First,
239 2nd or 3rd-generation immigrants frequently visit family in their country of origin[24], increasing
240 the odds of contact with MTBC strains endemic in Northern Africa. Second, migration waves
241 from Northern Africa began before those from Asia and sub-Saharan Africa. Compared to other

migrants, patients born in Northern Africa were older (median age 55y vs. 33y, $P < 10^{-6}$, Mann-Whitney *U*-test) and tended to have stayed in France longer before tuberculosis was diagnosed (median 13y vs. 4y, $P = 0.003$), increasing the odds of being infected with a French-endemic strain. This situation has probably contributed to blur the distinction between native and Northern African patients. It is unlikely, however, to have biased our conclusion regarding the major contribution of contacts with Northern Africa to the population structure of MTBC in France.

To conclude, we show that tuberculosis genotyping data obtained from a routine surveillance programme convey enough phylogenetic information to extract meaningful inferences regarding past epidemiological events. Our approach, based on the easily implemented THD technique, allowed to consider MTBC as a whole rather than to focus on a single lineage, as was the case with most recent phylogeographical studies of tuberculosis. The application of similar approaches to international settings such as Europe, which combines a low tuberculosis prevalence and intense migration flows, might provide important information relevant to the current re-emergence of tuberculosis.

Methods

Ethics statement. This retrospective, cross-sectional, observational multicentric study was approved by the Comité de Protection des Personnes Sud-Est IV under no. DC-2011-1306. Written consent of participants was not obtained, in accordance with French regulations, due to anymous treatment of data and the non-interventional nature of the study.

Patient population and collection of data. The 2008-2014 tuberculosis patient cohort of the ORAM was described previously and detailed data were made publicly available[18]. Briefly, 1,746 patients with available MTBC strains were identified. Countries of birth were coded according to ISO3166 standard (http://www.iso.org/iso/home/standards/country_codes.htm) and assigned to world regions according to the United Nations Geoscheme (http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm) with the exception of Central, Eastern, Southern and Southeastern Asia UN regions (n = 1, 4, 9 and 28 patients, respectively), which were pooled into a single Asia region to avoid small per-region sample sizes. Of note, we also complied with the UN Geoscheme convention that the Russian Federation, despite its territory spanning both Europe and Asia, was assigned to the Eastern Europe world region. Exclusion criteria were: (i), ambiguous MIRU-VNTR profile (i.e., undefined number of repeats at any of the 15 loci; n = 105); (ii), unknown country of origin (n = 958); (iii), birth outside Europe, Africa and Asia due to small sample size (n = 6); and (iv), infection with species whose transmission does not primarily involve interhuman contact (n = 39), including *M. bovis*, *pinnipedii* and *microti*, to avoid interpretation bias.

277 **Isolate genotyping and family/lineage assignment.** Spoligotyping was performed as described
278 elsewhere[16]. Spoligotypes were compared to those of the SpolDB4 database[25] to assign isolates
279 to families including AFRI, Beijing, BOV, Cameroon, CAS, Haarlem, LAM, S, T and X[26], which were
280 then reclassified into 6 major genome sequence- (or genomic deletion-) based lineages,
281 including e.g. the East-African Indian, East Asian, Euro-American, Indo-Oceanic and West African
282 lineages[27], according to known correspondences[28].

283 **Timescaled haplotypic density.** THD was implemented for the R platform (The R Foundation for
284 Statistical Computing, Vienna, Austria) using publicly available software code[18]. Briefly, THD
285 assigns a measure a density to a haplotype (here, a MIRU-VNTR genotype) based on its position
286 in the metric space defined by the genetic distances between a set of haplotypes. The density
287 function is estimated using kernel density estimation with a geometric kernel, where the kernel
288 bandwidth parameter is expressed in units of time (here, the timescale) based on a functional
289 relationship between genetic distance and the time to the most recent common ancestor
290 (TMRCA) under the infinitely many alleles model[29]. In its original implementation, THD was
291 measured for a given haplotype relative to all other haplotypes in a population, and interpreted
292 as an estimator of the number of common ancestors during the considered timescale, a
293 quantity related to the transmission success of the pathogen haplotype. In the present study,
294 THD was measured relative to haplotypes in other groups to reflect the intensity of transmission
295 events between groups during the timescale (rather than within a single group). Cross-THDs
296 between groups based on geographic regions were reported as the means of individual THDs in
297 each group, normalized to add to 1. Cross-THD from a group relative to itself was referred to as

298 self-THD. THD parameters, namely timescales and mutation rate, were set as described in our
299 previous study of MIRU-VNTR data[18].

300 **Multidimensional scaling.** Cross-THDs were transformed into dissimilarity (distance) measures
301 to ease visualization by means of multidimensional scaling[30]. The transformation involved: (i)
302 adding the asymmetric matrix of inverse pairwise cross-THDs to its transposed matrix to obtain
303 a symmetric matrix; (ii) normalizing the matrix by dividing each element with the outer product
304 of the square roots of its diagonal entries, similar to the normalization of a covariance matrix to
305 a correlation matrix, to obtain a matrix with unit diagonal elements; and (iii) subtracting 1 to all
306 elements of the normalized matrix so that the distance from one point to itself is zero. It is
307 easily verified that the elements of the resulting matrix fulfill the three conditions of a distance
308 measure (formally, a semimetric), namely non-negativity, symmetry and identity of the
309 indiscernibles. Finally, the THD-based dissimilarity matrix was taken as input to the
310 multidimensional scaling procedure, which projected each group as a point on a plane such that
311 inter-group distances in the plane were the best linear approximations of THD-based
312 dissimilarity.

Acknowledgements

We are indebted to all the microbiologists, physicians and technical staff who contributed to the ORAM, transmitted information and referred MTBC strains to reference laboratories for typing. We also thank E. Rivollier (ULAT of Saint-Etienne, France), T. Ferry, T. Perpoint and F. Valour (Lyon University Hospital) for help with data collection, and S. Mona, S. Boitard, M. Veuille (Muséum National d'Histoire Naturelle, France) and S. Nieman (Research Center Borstel, Germany) for fruitful discussion.

Author Contributions

JPR designed the study; OD, CP, GC, ASRB, GB, CGB, SB, AC, VJ, IF, MPM, CG and FA collected data; MB, TW and JPR performed the experiments and analyzed the data; JPF and GL contributed reagents/tools; MB, PS, TW and JPR wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing financial interests: The authors declare no competing financial interests.

References

1. Couvin, D. & Rastogi, N. Tuberculosis - A global emergency: Tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. *Tuberculosis (Edinb)* **95 Suppl 1,** S177-189 (2015).
2. The Lancet Respiratory Medicine. Changing minds about tuberculosis. *Lancet Respir Med* **3,** 901 (2015).
3. WHO | Global tuberculosis report 2015. *WHO* Available at: http://www.who.int/tb/publications/global_report/en/. (Accessed: 23rd February 2016)
4. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45,** 1176–1182 (2013).
5. Barbier, M. & Wirth, T. The Evolutionary History, Demography, and Spread of the *Mycobacterium tuberculosis* Complex. *Microbiol Spectr* **4,** (2016).
6. Wirth, T., Meyer, A. & Achtman, M. Deciphering host migrations and origins by means of their microbes. *Mol. Ecol.* **14,** 3289–3306 (2005).
7. Linz, B. *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445,** 915–918 (2007).
8. Moodley, Y. *et al.* Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.* **8,** e1002693 (2012).
9. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4,** e1000160 (2008).
10. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48,** 1535–1543 (2016).

11. Reed, M. B. *et al.* Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J. Clin. Microbiol.* **47,** 1119–1128 (2009).
12. Dale, J. W. *et al.* Origins and properties of *Mycobacterium tuberculosis* isolates in London. *J. Med. Microbiol.* **54,** 575–582 (2005).
13. Pichat, C. *et al.* Combined Genotypic, Phylogenetic, and Epidemiologic Analyses of *Mycobacterium tuberculosis* Genetic Diversity in the Rhône Alpes Region, France. *PLoS ONE* **11,** e0153580 (2016).
14. Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47,** 242–249 (2015).
15. Fallico, L. *et al.* Four year longitudinal study of *Mycobacterium tuberculosis* complex isolates in a region of North-Eastern Italy. *Infect. Genet. Evol.* **26,** 58–64 (2014).
16. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35,** 907–914 (1997).
17. Supply, P. *et al.* Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **44,** 4498–4510 (2006).
18. Rasigade, J.-P. *et al.* Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis. *Sci Rep* **7,** 45326 (2017).
19. Supply, P. *et al.* Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.* **39,** 3563–3571 (2001).

20. Eckart, W. U. *Man, Medicine, and the State: The Human Body as an Object of Government Sponsored Medical Research in the 20th Century*. (Franz Steiner Verlag, 2006).
21. Institut National de la Statistique et des Etudes Economiques. Répartition des étrangers par nationalité en 2014. (2014). Available at: https://www.insee.fr/fr/statistiques/2381750. (Accessed: 13th July 2017)
22. D'Albis, H. & Boubtane, E. Caractérisation des flux migratoires en France à partir des statistiques de délivrance de titres de séjour (1998-2013). *Population* **70,** 487–523 (2015).
23. Khyatti, M. *et al.* Infectious diseases in North Africa and North African immigrants to Europe. *Eur J Public Health* **24 Suppl 1,** 47–56 (2014).
24. Armand, L. Les touristes français à l'étranger en 2006: résultats issus du suivi de la demande touristique (Direction du Tourisme). *Bulletin Epidémiologique Hebdomadaire.* **25,** 218–221 (2007).
25. Brudey, K. *et al. Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* **6,** 23 (2006).
26. Comas, I., Homolka, S., Niemann, S. & Gagneux, S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE* **4,** e7815 (2009).
27. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 2869–2873 (2006).
28. Shabbeer, A. *et al.* TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect. Genet. Evol.* **12,** 789–797 (2012).

29. Walsh, B. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* **158,** 897–912 (2001).

30. Izenman, A. J. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. (Springer Publishing Company, Incorporated, 2008).

Conclusions et perspectives

Au cours de cette thèse j'ai utilisé des outils de génétique des populations et de phylogénie afin d'élucider certaines parties de l'histoire évolutive de la tuberculose, dans différents contextes épidémiologiques et grâce à différents types de données génétiques. Nous avons pour cela réuni des centaines de génomes complets et des données génotypiques de plusieurs milliers de souches. Nous avons analysé les séquences grâce à différentes méthodes phylogénétiques, des inférences Bayésiennes et divers tests statistiques.

Taux de mutation, origine, co-évolution avec l'homme

Dans le chapitre 1 nous avons fait une revue des connaissances actuelles sur le MTBC, rappelant sa structure phylogénétique et les lignées qui le compose, son origine, son histoire démographique et sa coévolution avec l'Homme. Par ailleurs nous avons mis en évidence certains points encore controversés, comme le taux de mutation et subséquemment l'âge estimé du MTBC. Depuis la publication du papier de Bos et al. 2014, deux visions concernant l'âge de la tuberculose se confrontent l'une à l'autre. La première, développée dans Comas et al. 2013 estime l'âge de la tuberculose comme étant proche de celui de l'homme moderne, précédant l'Out-of-Africa, à 70000 ans. La seconde approche estime l'apparition de la tuberculose comme étant bien plus récente, il y a seulement 6000 ans. L'extrême différence entre ces deux estimations s'explique par les méthodes employées. Dans l'étude de Comas *et al.*, les auteurs faisaient l'hypothèse à priori d'une coévolution de longue date entre l'homme et *M. tuberculosis*, se basant sur la comparaison des phylogénies des deux espèces qui présentent de fortes similarités et sur l'observation d'une forte association de certaines lignées du MTBC à des populations humaines (Gagneux 2012). Ils ont donc choisi un point de calibration dans l'arbre phylogénétique de la tuberculose en encrant certains clades sur des dates connues de l'histoire humaine en faisant effet miroir sur la phylogénie mitochondriale. Dans le cas présent, le point de coalescence des lignées 5 et 6 de la tuberculose avec le temps de coalescence de l'haplotype mitochondrial L3 de l'homme. Le taux de mutation ainsi estimé est de 2.58×10^{-9} substitutions par site et par année. Dans la seconde étude, les auteurs ont séquencé des génomes bactériens provenant de lésions tuberculeuses retrouvées sur des momies péruviennes datant de 1000 ans. Ces souches anciennes se sont révélées être des souches proches de *M. pinnipedii* actuelles. En se basant sur une gamme de paléogénomes disponibles ils ont utilisé la méthode dites de tip-dating afin d'estimer le taux de mutation à partir de l'âge des séquences. Le taux de mutation inféré est de 7.07×10^{-8} substitutions par

site et par année, datant l'origine du MTBC à 5268 avant JC. La méthode utilisée peut être contestable et certaines limitations existent: génomes anciens enclins à avoir des mutations causées par la dégradation environnementale de l'ADN, réactions d'hydrolyse, dégradation préférentielle des pyrimidines dans l'ADNa et transformation en de nouvelles bases chimiquement modifiées, utilisation des seules substitutions synonymes ayant souvent un taux de mutation supérieur aux mutations non synonymes soumises à des pressions de sélection plus importante et possibilité de variations (relaxation) du taux de mutations dans les lignées animales causées par le changement d'hôte. Avec une origine aussi tardive de la tuberculose, les similarités phylogéographiques du MTBC et d'*Homo sapiens,* ainsi que tous les indices indiquant une présence de la tuberculose avant cette date seraient uniquement le fruit du hasard, voir erronés. Ce scénario semble peu probable. Cependant, en considérant que les points méthodologiques cités plus haut sont négligeables, cette estimation du taux de mutation de *M. tuberculosis* n'est pas incompatible avec l'hypothèse de Comas et *al.* En effet, lorsque nous avons estimé le taux de mutation épidémique de bacilles de la tuberculose comme dans les chapitres 2 et 3 de cette thèse, en implémentant la méthode de tip-dating avec des souches contemporaines, nous obtenons un taux de mutation encore plus élevé, d'environ 1×10^{-7} substitution par site et par année. Si l'on estimait l'âge du MTBC avec ce taux, nous trouverions une date d'apparition encore plus récente. Ce hiatus peut être expliqué par le fait que le taux de mutation estimé varie selon l'échelle de temps considérée (Ho et al. 2011). En effet, en prenant en compte uniquement des souches récentes, le taux de mutation estimé sera défini comme spontané ou épidémique. Mais au fils du temps, sous l'effet de la sélection, beaucoup de mutations délétères observées sur ces souches disparaitront sur le long-terme. Ainsi, plus l'échelle de temps considérée sera grande et plus le taux de mutation inféré sera faible. On parlera alors de taux de mutation évolutif ou taux de fixation. D'autres facteurs, de type démographique ou des événements de spéciation, peuvent également influencer cette estimation. En définitive, le taux estimé dans *Bos et al.* est inférieur au taux de mutation épidémique estimé uniquement à partir de souches récentes mais encore supérieur au taux de fixation, qui serait celui estimé par Comas *et al*. Pour résoudre ce cas, il faudra donc découvrir des souches plus anciennes encore et réussir à séquencer le génome de ces dernières. Si cette hypothèse est juste, plus les souches utilisées seront anciennes et plus les taux de mutations inférés seront bas.

En attendant, ce problème épineux persiste dans les inférences effectuées sur la tuberculose. On trouve un exemple parlant en comparant deux publications : (Luo et al. 2015; Merker et al.

2015). Dans les deux études, des souches de la lignée Beijing sont utilisées afin d'inférer leur démographie. Dans celle de Merker *et al*., le taux de mutation est calculé par tip-dating tandis que dans Luo *et al*., le taux de mutation utilisé est celui estimé par Comas et *al.* Les fluctuations démographiques du skyline Bayésien sont logiquement similaires dans les deux études, cependant l'échelle de temps considérée et donc les dates des différentes oscillations populationnelles sont totalement différentes, changeant presque d'un facteur 100. Il est difficile pour les raisons citées plus haut de déterminer le scénario réel parmi ces deux inférences ; la réalité se trouvant vraisemblablement à mi-chemin. Ceci dit, la distorsion est d'autant plus importante que l'on s'éloigne du point temporel à partir duquel le taux de substitution a été calibré. Une perspective pour les futurs travaux dans ce domaine serait d'utiliser un taux de mutation exponentiellement décroissant décrivant le passage du taux de mutation spontané au taux de fixation au cours du temps. Les paramètres de cette décroissance pourraient être déterminés en utilisant de nouveaux génomes anciens, avec une bonne couverture temporelle. L'approche alternative consiste à étendre la stratégie de Comas *et al.* en effectuant un calibrage de la phylogénie du MTBC, mais en utilisant davantage de points d'ancrage. Cependant dans le cadre d'inférences démographiques Bayésiennes, comme sous BEAST (Bouckaert et al. 2014), cela demeure un problème ardu car la population efficace est directement estimée à partir du taux de mutation ($\theta = 2N_0\mu$). Ainsi le modifier entraînerait dans le même temps une modification de la taille efficace estimée. L'implémentation d'un taux de mutation fluctuant dans un modèle démographique tel que le skyline Bayésien parait donc difficile.

Impact des mutations compensatoires

Dans le second chapitre, nous nous sommes intéressés à un problème de santé public émergent et particulièrement préoccupant, les épidémies de souches MDR. Pour cela nous avons analysé les génomes complets de souches isolées dans des contextes épidémiologiques précis. Un premier jeu de données composé de 277 souches provenant du Karakalpakstan collectées entre 2001 et 2006 et un second de 428 souches prélevées entre 2008 et 2010 dans la région de Samara en Russie. Un clade particulier, appartenant à la lignée Beijing, représentait la majeure partie de ces souches résistantes. En nous concentrant sur ces dernières nous avons mis en évidence des mutations dites compensatoires ayant pu favoriser l'accumulation et la maintenance de hauts niveaux de résistance dans ces souches. Nous avons également démontré l'impact de l'implémentation des DOTS dans l'établissement des souches MDR en Asie centrale.

Dans le but de conforter ces résultats, nous nous sommes penchés sur un autre clade de la lignée Beijing, proche du précédent, le clone W148, presque exclusivement composé de souches MDR. Nous nous sommes intéressés à son origine, sa démographie et sa dispersion à l'échelle continentale. Comme dans le cas précédent, la présence de mutations compensatoires dans les gènes *rpoC*, *rpoA* ou *rpoB*, a confirmé un net avantage sélectif aux souches les portant, se traduisant notamment par l'accumulation de mutations conférant des résistances aux antibiotiques d'une part, mais aussi par une meilleure transmissibilité des souches hautement résistantes d'autre part. Ces trois gènes ciblés sont porteurs de mutations compensatoires conférant une meilleure résistance à la rifampicine (Li et al. 2016; De Vos et al. 2013; Brandis et al. 2012). Enfin, nous avons pu montrer l'impact considérable de certaines de ces mutations sur la fitness et la transmissibilité des souches résistantes. Théoriquement des mutations compensant l'acquisition de résistances à d'autres antibiotiques doivent exister. Cela pourrait être un autre axe de recherche majeur, tant la découverte de tels gènes pourrait aider à comprendre le succès de certaines souches hautement résistantes. Le rôle des mutations compensatoires dans l'évolution de souches résistantes vers des profils MDR ou XDR n'est que partiellement compris, des progrès en ce sens permettraient d'adapter les traitements et de soigner au mieux les patients.

Par inférences Bayésiennes nous avons déterminé qu'au cours du dernier siècle, cette lignée originaire d'Asie centrale a disséminé via la Russie jusqu'en Europe de l'ouest et a vu sa population être multipliée par un facteur 100, possiblement favorisée par des mutations au sein de gènes impliqués dans l'expression de protéines antigéniques, dans la dormance et dans les systèmes de réparation de l'ADN. En effet, sur 74 mutations détectées, 16 affectaient des gènes impliqués dans la réponse immunitaire de l'hôte ou présentant des épitopes membranaires, dont 6 étaient des mutations non synonymes. Ces faits sont intéressants car l'origine de ce clade a été inférée en Asie centrale, et si l'on remonte plus loin encore, appartenant à la lignée Beijing, celui-ci vient originellement d'Asie de l'Est. A l'heure actuelle ce clade est particulièrement prévalent en Russie, dans les pays de l'ex URSS et en Europe de l'Est. La lignée Beijing quant à elle, bien que majoritairement présente en Asie (44.7%), montre une prévalence importante en Europe également (27.9%) (Ramazanzadeh and Sayhemiri 2014). Pourtant ces souches, au même titre que le CAO (Central Asian Outbreak) évoqué lors du chapitre 2 n'auraient commencé à se disperser en Europe qu'au début du siècle dernier. Une explication à leur succès, en lien avec les mutations touchant à la réponse antigénique humaine, pourraient être que ces lignées, coévoluant avec des

populations asiatiques depuis des milliers d'années, ont découvert un hôte « naïf » ou plutôt moins bien adapté dans les populations indo-européennes. En effet, il a été montré que certaines lignées du MTBC étaient associées à des populations humaines (Fenner et al. 2013; Asante-Poku et al. 2015; Gagneux et al. 2006). Un contexte nouveau ou une rupture d'équilibre expliquerait le caractère épidémique de ces souches et la mise en place récentes de mutations adaptatives. De la même manière, lors de la colonisation et des grandes explorations, les souches de la Lignée 4, indo-européenne, dispersées par les populations européennes ont fait des ravages, notamment en Afrique et en Amérique au sein de populations naïves vis-à-vis de cette lignée (Bates and Stead 1993). Ces éléments de réponse semblent démontrer la longue coévolution liant l'homme et la tuberculose, ainsi que ses radiations secondaires. Dans le cas où le pathogène rencontrerait un individu issu d'une population différente, celui-ci pourrait transitoirement être plus virulent car le système hôte pathogène ne serait pas à l'équilibre. Avec le développement exponentiel des capacités de séquençage mais également des outils d'analyses, le futur réside indéniablement dans l'analyse conjointe de génomes complets de l'hôte (*Homo sapiens*) et de son pathogène (*M. tuberculosis*), permettant ainsi d'étudier plus finement les interactions hôtes pathogènes et d'évaluer les associations et susceptibilités entre populations humaines et lignées du MTBC, amenant un jour peut-être à une médecine personnalisée pour chaque patient souffrant de la tuberculose. Ainsi, des projets croisant des cohortes de patients atteint de tuberculose (contrôle positif) d'une même population et des cohortes de personnes non-atteintes (test à la tuberculine cutanée négatif) permettraient éventuellement de détecter par des méthodes d'associations génotype-phénotype à échelle génomique (GWAS) des gènes/allèles qui corrèlent avec le développement d'une infection pathologique. Mieux encore, on peut songer à croiser les données de puces à ADN permettant d'analyser jusqu'à 2,5 millions de polymorphismes (SNP's) de patients humains avec les génomes *de M. tuberculosis*, en intégrant la dimension taxonomique et géographique de l'agent pathogène. Ce type d'étude est envisageable dans des pays qui permettent l'accès à un échantillon de sang de toute la population; comme c'est le cas notamment au Danemark.

Nous avons par la même occasion identifié la présence de mutations fixées au sein du clone W148 dans des gènes impliqués dans la dormance, l'acquisition du fer et la réparation de l'ADN, toutes des fonctions essentielles dans la pathogénicité et la persistance de *M. tuberculosis* dans son hôte. Le simple listing de ces gènes est d'un intérêt tout relatif, l'intérêt adaptatif des SNP's identifiés gagnerait à être confirmé en pratiquant des expérimentations

sur des modèles murins ou macrophages en infectant ces derniers avec des souches identiques possédant un allèle sauvage ou leur variant d'intérêt, et ce, dans la veine de la publication de Gonzalo-Asensio et al. 2014. Dans le but d'évaluer l'importance de la fonction d'un gène il est également important de le considérer dans la cadre plus large de son réseau métabolique, en quittant une vision nucléaire du gène pour s'accaparer celle de son interactome. Ainsi, dans la cadre de nos tests de sélection, l'émergence d'une succession de SNPs touchant des gènes d'une même voie métabolique, renforcerait la dimension statistique du signal et de la cible. Malheureusement nos connaissances demeurent limitées et une large fraction des gènes qui composent le génome de *M. tuberculosis* n'ont toujours pas de fonction connue ; laissant de nombreuses pièces des puzzles métaboliques de côté.

Par ailleurs, il est important de noter qu'une famille entière de gènes reste encore très peu connue. Cette famille de gènes appelés PE/PPE, compte pour une très grande part du génome, approximativement 7%, comprenant environ 100 gènes PE et 70 gènes PPE, mais elle est très compliquée à étudier de par sa dimension répétitive et ses motifs récurrents. Ces gènes sont uniques aux mycobactéries et encore plus abondants au sein de mycobactéries pathogènes dont fait partie la *M. tuberculosis* (Sampson 2011). Leur particularité est la présence de motifs Proline-Glutamate (PE) et Proline-Proline-Glutamate (PPE) dans l'extrémité N-terminale de la protéine exprimée. Cette extrémité N-terminale est relativement bien conservée dans cette famille et fait entre 100 et 200 acides aminés. Le reste des gènes PE-PPE, au contraire, est extrêmement variable contenant de longues séquences riches en GC. Ainsi, certains gènes codent pour des protéines composées uniquement de la partie N-terminale tandis que d'autres atteignent une longueur de 3700 acides aminés (Mukhopadhyay and Balaji 2011). Ces gènes étant donc extrêmement homologues et répétés, les techniques de séquençage NGS les plus usités, utilisant de nombreuses séquences courtes (entre 100 et 300 paires de bases) comme les technologies Illumina, ne permettent pas de les séquencer précisément. En conséquence, elles sont souvent exclues des études génomiques consacrées à *M. tuberculosis*. Pourtant ces gènes sont d'un intérêt crucial, soupçonnés d'être le siège de pressions de sélection positive, d'évènements de recombinaisons et dont les protéines sont impliquées dans la réponse antigénique humaine (Phelan et al. 2016). Certains de ces protéines interagissent avec TLR2, récepteur de type Toll et servant à la reconnaissance bactérienne chez l'homme, induisant la sécrétion de cytokines par les macrophages et les cellules dendritiques et promouvant l'apoptose chez les cellules hôtes. D'autres gènes PE/PPE semblent liés au système de sécrétion ESX5 de type VII et jouent un rôle dans la stabilité de l'enveloppe extérieure du

bacille (Brennan 2017). Un autre élément extrêmement conservé, PE-PGRS 62, possédant un rôle dans la réplication et la persistance de la bactérie a été proposé comme un bon candidat pour le développement d'un nouveau vaccin antituberculeux. En conclusion, ces gènes semblent avoir un grand nombre de fonctions, antigénique entre autres, et montrent une grande variabilité, évoluant sous l'action de la sélection positive et de la recombinaison. Leur implémentation dans les études en génome complet et un approfondissement de nos connaissances sur leur activité et fonction semble essentielle pour le développement de nouveaux traitements. L'essor de nouvelles techniques de séquençage et des coûts décroissants, permettront de démocratiser l'usage de technologie de séquençage SMRT par PacBio permettant de produire des séquences atteignant jusqu'à 10000 paires de bases (Buermans and den Dunnen 2014), permettant ainsi de contourner les erreurs d'assemblage inhérentes à l'Illumina.

Reconstructions démographiques

Dans les chapitres 2 et 3, nous avons estimé les changements de taille efficace de plusieurs lignées du MTBC. Pour ce faire, nous avons utilisé le programme d'inférences Bayésiennes de phylogénies BEAST, en comparant notamment deux modèles démographiques. Un premier modèle de taille de population constante et un second, le skyline Bayésien (Drummond et al. 2005), autorisant la taille effective de la population à varier. Le modèle de skyline Bayésien était favorisé dans les deux études et indiquait une forte hausse de la taille efficace de la population au cours du temps. Cependant, ce point fait émerger des questionnements. Premièrement que représente la taille efficace de la population dans le cadre de bactéries, et plus précisément de la tuberculose? La taille efficace d'une population est censée représenter le nombre d'individus dans une population idéale, évoluant uniquement sous l'action de la dérive génétique avec accouplement aléatoire, pour montrer la même diversité génétique que la population idéale. Dans le cadre d'une bactérie haploïde se multipliant par scission, cette statistique peut toujours être calculée mais il est difficile d'évaluer ce qu'elle représente réellement. Evalue-t-on le nombre de personnes infectées, le nombre de bactéries (un unique patient infecté peut porter une charge bactérienne de plusieurs millions de bacilles) ou le nombre de souches transmises par aérosols d'une personne à l'autre ? Ceci n'est pas une difficulté insurmontable car lors des inférences démographiques la valeur de la population efficace n'est pas cardinale, ce sont les fluctuations de cette taille efficace qui sont étudiées. Cependant un second problème se pose ici, à savoir la pertinence de ces estimations dans le cas de la tuberculose. En effet, les variations de la taille efficace

sont vues comme des changements de tailles démographiques sous l'hypothèse du modèle standard neutre, c'est-à-dire dans une population non structurée, échantillonnée au hasard et sans recombinaison et sélection. Or dans le cas de *M. tuberculosis* ces conditions sont pour la plupart violées, principalement celle de structuration de population. Les populations du bacille de Koch évoluent par scission, sans recombinaison, et forment des entités génétiquement et spatialement structurées. Or ces écarts au modèle standard neutre peuvent provoquer des variations de la taille efficace sans changements démographiques de la population (Lapierre et al. 2016). Bien que largement utilisées, au cours de cette thèse également, les inférences démographiques dans le cadre de la tuberculose doivent être interprétées avec prudence ; mais ces questionnements ne sont pas spécifiques à notre système biologique.

Souches animales

Pour continuer d'étudier les adaptations possibles du pathogène de la tuberculose au cours de son évolution, nous sommes retournés à une échelle plus large et nous sommes intéressés aux grandes lignées composant le MTBC et notamment les lignées animales, bien moins connues que les souches infectant l'homme. Nous avons collecté et analysé les génomes complets d'un grand nombre de souches animales combiné à des représentants de chaque lignée humaine. Les phylogénies obtenues semblent en accord avec les connaissances actuelles, séparant un embranchement purement humain comprenant les lignées 1, 2, 3, 4 et 7, d'un embranchement mixte composé de lignées humaines et animales. Cependant, l'enracinement en midpoint rooting intercale la lignée 5 comme étant un groupe sœur des lignées humaines modernes. Nous avons par la suite mis en place un protocole destiné à détecter des mutations pouvant expliquer l'adaptation de ces souches à leurs différents hôtes au travers de nouveaux traits adaptatifs liés à des mutations ponctuelles sur certains gènes. Lorsque nous aurons dressé une liste de gènes d'intérêts grâce à ces méthodes, nous les testerons pour la sélection en utilisant le package PAML permettant de tester différents modèles de sélection sur des séquences codantes. Le risque de cette méthode, dans le cas de la tuberculose, est que le nombre de variations par gènes soit trop faible pour avoir un support statistique suffisant ; une approche pour contourner cette limitation est de travailler par catégories de gène ou sur une famille de gènes particulièrement pertinente.

Depuis que la transmission de la tuberculose de l'homme à l'animal semble acquise, une hypothèse forte de ces transferts intergénériques est que ces transmissions ont connu un apex il y a environ 10000 ans lors de la domestication (Smith et al. 2009, Wirth el al 2006).

Cependant quelques faits tendent à faire penser que la tuberculose aurait infecté les animaux depuis plus longtemps. Premièrement les traces d'une infection par la tuberculose ont été retrouvées sur les restes d'un bison datant de 17000 ans (Lee et al. 2015). Vraisemblablement si ce bison était bien infecté par la tuberculose, il y a de grandes chances que ce soit par *M. bovis*. Comme nous l'avons vu lors de cette étude, les souches animales font toutes parties d'un même clade relativement touffu, plus ou moins intriquées avec la lignée humaine 6. Cette dernière lignée est une lignée humaine confinée à l'Afrique de l'Ouest. Or si le passage de l'homme à l'animal s'était fait durant la domestication, ayant principalement eu lieu dans le croissant fertile et en Asie de l'est, on s'attendrait logiquement à ce que la lignée animale ait un ancêtre commun proche des lignées modernes du MTBC (Lignées 2, 3 et 4). Selon l'hypothèse que j'avance, la lignée animale trouverait son origine dans des temps bien plus anciens que la domestication. Afin d'aller plus loin dans ces investigations, nous prévoyons, d'une manière similaire à Comas et al. 2013, de tester différents scénarios d'origine évolutive de la lignée animale en utilisant différents points et temps de calibration.

Enfin, il est probable que dans le cadre d'une étude sur les souches animales, utiliser la souche H37rv, souche de la lignée 4, comme référence lors de l'assemblage ne soit pas la méthode optimale. Ainsi nous prévoyons, d'effectuer des analyses similaires mais avec un jeu de donnée obtenue en utilisant une souche de référence de *M. bovis* pour l'assemblage. Le fait d'utiliser une souche animale, bien plus proche donc, en référence, pourrait modifier, au moins légèrement, les SNPs identifiés. On peut s'attendre à obtenir plus de mutations inter souches animales et moins entre souches des lignées humaines.

Des analyses complémentaires peuvent être réalisées, notamment en assemblant de novo les génomes des souches de notre collection afin de s'affranchir de biais possibles dus au choix de la souche de référence pour l'assemblage. Cependant cette méthode est beaucoup plus intensive et chronophage informatiquement parlant, et peut être limitée par la taille des fragments obtenus par séquençage Illumina comme nous l'avons évoqué dans le cadre des familles de gènes PE/PPE. Cette approche permettrait également de mieux appréhender les potentiels événements de délétions et acquisitions de gènes le long des lignées animales. Ce type d'information peut-être d'un intérêt majeur, puisque les gènes sous forte relaxation de pression de sélection ou perdus dans les souches animales sont de très bons candidats pour représenter des gènes essentiels à la virulence chez l'homme où à l'interface du système immunitaire pathogène/hôte.

THD

Lors des deux derniers chapitres nous nous sommes intéressés à la notion de succès d'une souche dans un contexte épidémiologique. En effet, lors des chapitres 2 et 3 nous avons constaté la difficulté d'évaluer de façon satisfaisante cette dernière. Or ce paramètre peut être particulièrement utile afin d'évaluer les facteurs qui contribuent à la transmissibilité des souches. Nous avons donc développé une méthode, que nous avons appelé THD (Timescaled haplotypic density), permettant d'estimer le succès individuel d'une souche. Nous l'avons défini comme la fréquence des événements de transmissions auxquelles elle est associée au cours d'une certaine période, avec la possibilité de moduler l'échelle de temps considérée. Pour valider cette méthode nous avons analysé les génotypes MIRU d'isolats obtenus lors d'une étude de surveillance d'une cohorte de 1641 patients en Rhône-Alpes. La méthode s'est révélée particulièrement pertinente lorsque nous avons associé le THD à des données sociodémographiques, concernant les patients touchés, ainsi que des informations relatives à la pathogénicité des souches et les symptômes provoquées. De plus, nous avons démontré l'intérêt du THD dans le cadre plus large de questions phylogéographiques en soulignant la similarité génétique des souches françaises avec celles d'Afrique du Nord et en démontrant la baisse de la contribution des souches en provenance d'Afrique subsaharienne et d'Europe de l'Est dans le paysage infectieux de la tuberculose en France. Ce nouvel indice épidémiologique, simple et efficace, est une première dans le cadre de l'étude de la tuberculose et permet d'évaluer quantitativement le succès des souches, qu'on ne pouvait auparavant évaluer que qualitativement à l'aide de réseaux de type Minimum Spanning Tree, de phylogénies ou sur la base de données épidémiologiques. Par ailleurs, le fait de pouvoir moduler l'intervalle de temps pris en compte permet de faire ressortir des informations plus difficilement accessibles et moins évidentes. Basé sur des distances, le THD permet d'analyser de gros jeux de données et théoriquement de tous types (VNTR, séquences nucléotidiques, génomes). Plus le support génétique étudié évoluera à un rythme rapide et plus l'intervalle de temps considéré pourra être contrôlé finement. Ainsi nous espérons pouvoir appliquer cette méthode à des séquences de génomes complets dans un futur proche. Dans le cas de génomes complets, la distance utilisée correspond à une méthode grossière d'estimation de la réelle distance évolutive entre les souches, que seul un modèle de substitution nucléotidique peut corriger afin de tenir compte des substitutions homoplastiques (Voir annexe méthode). L'utilisation du THD dans ces conditions doit donc être restreinte à des échelles de temps courtes et moyennes car sur une échelle de temps importante, le temps

estimé pourrait dévier grandement. *M. tuberculosis* était un très bon organisme pour tester le THD car c'est un modèle simple ne recombinant pas, mais le THD peut très bien être utilisé pour étudier la transmission d'autres espèces bactérienne, à condition qu'elles évoluent essentiellement par clonalité. *Salmonella typhi*, *Yersinia pestis* ou bien *Streptococcus pneumonae* (Buermans and den Dunnen 2014) sont des candidats plausibles. La faiblesse majeure du THD réside dans le fait qu'il est extrêmement dépendant de l'échantillonnage et doit donc être utilisé uniquement dans des études utilisant un échantillonnage rigoureux représentant soit une population de manière exhaustive soit une part représentative et non biaisée de cette population. Quand ces conditions sont remplies, il s'agit d'un outil puissant permettant de quantifier l'association entre le succès d'un pathogène à plus ou moins long terme et des informations relatives aux patients, au pathogène, aux types de symptômes, ainsi qu'à un contexte socio-économique. Enfin, au cours du dernier chapitre nous avons montré qu'il permettait de quantifier les échanges entre plusieurs régions et donc d'évaluer l'influence de zones géographiques sur la population d'une autre. Ainsi le THD n'est pas seulement un outil d'épidémiologie mais peut également contribuer à la compréhension des patrons phylogéographiques et migratoires.

Bibliographie

Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, Stucki D, Hattendorf J, Borrell S, Feldmann J, Danso E, Gagneux S. 2015. Mycobacterium africanum Is Associated with Patient Ethnicity in Ghana. *PLoS Negl Trop Dis* **9**.

Bates JH, Stead WW. 1993. The history of tuberculosis as a global epidemic. *Med Clin North Am* **77**: 1205–1217.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest S a., Bryant JM, Harris SR, Schuenemann VJ, et al. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **10**: 1–6.

Brandis G, Wrande M, Liljas L, Hughes D. 2012. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. *Mol Microbiol* **85**: 142–151.

Brennan MJ. 2017. The Enigmatic PE / PPE Multigene Family Vaccination. *Infect Immun* **85**: 1–8.

Buermans HPJ, den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications.

Biochim Biophys Acta - Mol Basis Dis **1842**: 1932–1941.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**: 1176–1182.

De Vos M, Müller B, Borrell S, Black PA, Van Helden PD, Warren RM, Gagneux S, Victor TC. 2013. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob Agents Chemother* **57**: 827–832.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**: 1185–1192.

Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, Helbling P, Fehr J, Gsponer T, Rieder HL, et al. 2013. HIV Infection Disrupts the Sympatric Host-Pathogen Relationship in Human Tuberculosis. *PLoS Genet* **9**.

Gagneux S. 2012. Host-pathogen coevolution in human tuberculosis. *Philos Trans R Soc B Biol Sci* **367**: 850–859.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, et al. 2006. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **103**: 2869–2873.

Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, Laval F, Daffé M, Martin C, Brosch R, et al. 2014. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A* **111**: 11491–11496.

Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Mol Ecol* **20**: 3087–3101.

Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. *Mol Biol Evol* **33**: 1711–1725.

Lee OY-C, Wu HHT, Besra GS, Rothschild BM, Spigelman M, Hershkovitz I, Bar-Gal GK, Donoghue HD, Minnikin DE. 2015. Lipid biomarkers provide evolutionary signposts for the oldest known cases of tuberculosis. *Tuberculosis* **95**: S127–S132.

Li QJ, Jiao WW, Yin QQ, Xu F, Li JQ, Sun L, Xiao J, Li YJ, Mokrousov I, Huang HR, et al. 2016. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant Mycobacterium tuberculosis Beijing genotype strains in China. *Antimicrob Agents Chemother* **60**: 2807–2812.

Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G, et al. 2015. Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. *Proc Natl Acad Sci* **112**: 8136–8141.

Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, Rüsch-Gerdes S, Mokrousov I, Aleksic E, et al. 2015. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage (Supplementary). *Nat Genet*.

Mukhopadhyay S, Balaji KN. 2011. The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis* **91**: 441–447.

Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC, Glynn JR, Crampin AC, Alves A, et al. 2016. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17**: 151.

Ramazanzadeh R, Sayhemiri K. 2014. Prevalence of Beijing family in Mycobacterium tuberculosis in world population: Systematic Review and Meta-Analysis. *Int J Mycobacteriology* **3**: 41–45.

Sampson SL. 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin Dev Immunol* **2011**.

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon S V. 2009. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol* **7**: 537–544.

Annexe méthodes

Les méthodes décrites dans cette section sont développées grandement dans Ziheng Yang 2014.

Phylogénies

Les arbres phylogénétiques sont une représentation de la généalogie, des liens de parentés, entre des séquences pouvant représenter des espèces, des individus ou des gènes. Au cours de cette thèse sur la tuberculose, nous avons produits des arbres phylogénétiques, construits à partir de SNPs (Single Nucleotide Polymorphisms) obtenus grâce à des méthodes de WGS (Whole Genome Sequencing), afin de connaitre l'apparentement entre différentes souches. De nombreuses méthodes de reconstruction phylogénétique existent, dans les études présentées ici nous en avons utilisé principalement deux. Ce sont des méthodes basées sur les caractères, dans ce cas des nucléotides, qui ont pour but de produire un arbre en accord avec les caractères de chaque site et chaque séquence. Plus précisément, elles se basent sur des modèles de substitution des nucléotides. Les méthodes de construction phylogénétique les plus simples se basent sur les distances calculées deux à deux entre les séquences d'un jeu de données. La matrice des distances ainsi calculée peut être convertie en un arbre grâce à un algorithme de clustering.

Modèles de substitution

La méthode de mesure la plus simple consiste à compter le nombre de sites où les nucléotides sont différents entre deux séquences. Dans le cas où les séquences sont peu distantes, le résultat est proche de la réalité, mais cela devient vite erroné lorsque les divergences deviennent plus grandes. En effet, pour un site variable, plusieurs substitutions ont pu avoir lieu entre les deux séquences. Pour les sites non variables des substitutions suivies de réversions ou alors les même substitutions parallèlement. On voit bien ici qu'on ne peut tenir compte des changements nucléotidiques cryptiques qu'avec des modèles probabilistiques. Dans cette optique, des modèles de substitution nucléotidique ont été développés dans lesquels chaque site de la séquence est considéré comme évoluant indépendamment des autres. Pour chaque site, une chaîne de Markov décrit les substitutions. Ces chaînes de Markov ont pour états les 4 nucléotides, A, C, G et T et n'ont pas de mémoire : leur changement d'état dépend uniquement de l'état présent et pas de celui précédent. Il existe de nombreux modèles de complexité diverses. Dans le plus simple, JC69 (Jukes and Cantor 1969), le taux de substitution λ entre les nucléotides est le même pour

chaque substitution possible d'un nucléotide *i* à un autre nucléotide *j.* Les modèles plus complexes prennent en compte des réalités biologiques et contiennent donc des paramètres supplémentaires. Par exemple, dans le modèle HKY85 (Hasegawa et al. 1985), les transitions (changements entre deux pyrimidines T et C ou entre deux purines A et G) prendront un taux α différent des transversions (passage d'une purine à une pyrimidine ou le contraire) qui auront un taux β. De plus, ce modèle prend également en compte le fait que les séquences peuvent avoir des compositions en bases inégales (pas 25% par type nucléotidique), comme dans le cas de la tuberculose qui a un fort taux en GC par exemple. D'autres paramètres peuvent être ajoutés aux modèles de substitutions, comme un taux de substitution variable selon les sites. En effet, le taux de mutation mais également les pressions de sélection peuvent être différentes selon les sites d'une même séquence. Pour cela on assume que le taux de substitution suit une distribution statistique, souvent une distribution gamma dans ce cas, et varie selon les valeurs de la distribution estimée. Une propriété importante des modèles présentés ci-dessus est qu'ils sont « time reversible », d'autres modèles non présentés ici ne le sont pas. C'est-à-dire que la probabilité que la chaîne de Markov passe d'un état *i* à un état *j* est la même que de passer de *j* à *i* lorsque *i* n'est pas égal à *j.* Cela implique que la probabilité qu'une séquence *a* soit ancestrale à une séquence *b* est la même que la séquence *b* soit ancestrale à la *a*. Les arbres produits selon ces modèles ne sont donc pas enracinés et la racine peut être placée n'importe où. On peut penser que les modèles les plus complexes, comportant le plus grand nombre de paramètres, sont les plus justes. Cependant dans le cas de séquences proches, les différences peuvent être minimes et le temps de calcul doit être pris en compte. Plus les séquences seront longues, nombreuses et divergentes plus les modèles complexes seront nécessaires pour faire des estimations correctes.

Inférence d'abres par Maximum Likelihood

La première méthode de phylogénie que nous allons présenter est celle de Maximum Likelihood. Elle utilise un critère, nommé le score de vraisemblance, afin de mesurer à quel point les arbres prédisent les données en se basant sur un modèle d'évolution. L'arbre avec le meilleur score, la plus grande vraisemblance, sera jugé comme celui s'approchant le plus de la réalité. La vraisemblance est la probabilité d'observer les données lorsque des paramètres sont connus. Dans ce cas les paramètres seront donnés par le modèle de substitution et la probabilité des mutations de chaque site calculée en fonction. La vraisemblance de l'arbre sera la somme des probabilités calculées pour chaque site étant donné la topologie de l'arbre et le modèle de substitution. Dans les faits, l'arbre nécessitant le moins de mutations pour

expliquer sa topologie aura la plus grande vraisemblance, comme dans la méthode de maximum parcimonie où l'arbre optimal sera celui requérant le moins de changements sur la séquence. Mais la maximum likelihood sera calculée pour un modèle de substitution, prenant en compte les probabilités des différentes substitutions nucléotidiques et les taux de substitution. Le problème de cette méthode est donc de calculer le score d'un arbre et de rechercher les arbres pour calculer leur score. En théorie, il faudrait donc calculer le score de tous les arbres possibles pour un jeu de donnée et sélectionner l'arbre qui possède le meilleur et qui donc explique le mieux les données. En réalité, à partir d'une dizaine de séquences, cette recherche exhaustive est impossible car le nombre d'arbres devient beaucoup trop grand. Pour résoudre ce problème les programmes utilisent des algorithmes de recherche heuristique. Il existe de nombreux algorithmes de recherche heuristique, ceux-ci peuvent être classés en deux catégories. La première regroupe les algorithmes de regroupement hiérarchique qui peuvent être d'agglomération ou de division. Par étapes de fusion ou de division des branches, l'algorithme va choisir l'arbre intermédiaire présentant le plus haut score jusqu'à l'arbre final. La seconde catégorie regroupe les algorithmes d'échanges de branches et de réarrangement de l'arbre qui comme leur nom l'indique font des petits changements dans l'arbre et décide de garder le changement ou non à partir du score de l'arbre. Lorsque ce processus est répété de nombreuses fois sans amélioration du score, la chaîne s'arrête. Les algorithmes de recherche heuristique peuvent implémenter ces deux types d'algorithmes, le premier type pour obtenir un arbre de départ puis la seconde afin de l'optimiser par exemple. Pour choisir entre deux modèles, différentes méthodes sont disponibles. Pour deux modèles imbriqués (un des deux modèles est un cas précis de l'autre, il est l'hypothèse nulle) il y a le LRT (Likelihood Ratio Test) qui permet de rejeter l'un des modèles par rapport à l'autre en comparant le résultat à une table de chi-deux. Lorsque deux modèles ne sont pas imbriqués il est arbitraire d'en désigner un comme hypothèse nulle. Dans ce cas deux scores peuvent être calculés pour comparer les modèles. Le AIC (Akaike information criterion) est un score qu'on calcule pour chaque modèle, avec $AIC = -2\ell + 2p$. Où ℓ est le log likelihood optimal du modèle et p le nombre de paramètres du modèle. Le modèle avec le plus faible AIC est préféré. Cependant comme on le voit dans l'équation, l'AIC tend à favoriser les modèles complexes, riches en paramètres, puisque pour un paramètre supplémentaire, il suffit que l'augmentation de likelihood soit supérieure à 1 pour que le modèle soit préféré. Un autre score pénalisant plus les modèles riches en paramètres peut donc être utilisé, le BIC (Bayesian information criterion), avec $BIC = -2\ell + p\log(n)$, avec n la taille de la séquence. De la même manière le modèle avec le BIC le plus faible sera préféré.

Une méthode pour estimer la robustesse d'un arbre et de ses nœuds est la méthode de bootstrap. C'est une méthode de ré-échantillonnage à partir du jeu de données afin d'obtenir une mesure du support statistique des nœuds. Pour cela, pour une séquence de *N* nucléotides, nous allons tirer au hasard n nucléotides avec remplacement et reconstruire une phylogénie à partir des nouvelles séquences. Lorsqu'un nœud présent dans l'arbre original est présent dans l'arbre issu du ré-échantillonnage, il prend la valeur de 1 sinon il prend la valeur de 0. L'opération sera répétée *x* fois jusqu'à obtenir un pourcentage pour chaque nœud exprimant leur robustesse. Pour des gros jeux de données, malgré de nombreuses méthodes d'optimisation, le temps de calcul avec la méthode de maximum likelihood est très important. D'autres méthodes ont donc été développées, moins chronophage en calcul et permettant de tester des hypothèses évolutives et des modèles plus complexes.

Inférence d'arbres par méthode bayésienne

La deuxième méthode que nous allons présenter est l'inférence bayésienne de phylogénie. Les statistiques bayésiennes s'opposent aux statistiques classiques (fréquentistes) dans le fait que la probabilité d'un événement représente la fréquence à laquelle il se produira sur un grand nombre de répétitions dans l'approche fréquentiste. Dans les statistiques bayésiennes la probabilité représente plus un degré de certitude. Ainsi, alors qu'en statistiques fréquentistes seules les variables aléatoires possèdent une distribution tandis que les paramètres sont fixes et inconnus, dans les statistiques bayésiennes les paramètres ont une distribution traduisant les incertitudes les concernant. Ainsi la distribution d'un paramètre avant analyse des données sera appelée « Prior distribution » et la distribution du paramètre après analyse des données sera appelée « Posterior distribution ». Celle-ci combine les informations données par le prior et par les données. Dans le cadre de l'inférence bayésienne en phylogénie, la probabilité d'un arbre sera combinée à la likelihood des données afin d'avoir la probabilité postérieure de l'arbre. La probabilité postérieure de l'arbre représente la probabilité que l'arbre soit correct et donc l'arbre avec la plus grande probabilité postérieure sera choisi. Comme dans la méthode de maximum likelihood, l'estimation de la phylogénie se fait en utilisant les caractères et avec un modèle d'évolution sous-jacent. Cependant l'inférence bayésienne est plus efficace informatiquement (en temps de calcul par rapport à la méthode de bootstrap) pour évaluer la robustesse de l'arbre, permettant d'évaluer l'incertitude des branches en donnant la probabilité postérieure des clades, et permet également d'incorporer des modèles d'évolution plus complexes afin de tester différents scénarios évolutifs. Ces inférences ont été rendues possibles par le développement de l'algorithme

MCMC (Markov Chain Monte Carlo). De manière simplifiée, cet algorithme fonctionne de la façon suivante : 1) Un arbre est sélectionné aléatoirement. 2) Dans la sélection des possibles un arbre proche est sélectionné. 3) Un ratio des probabilités des deux arbres est calculé. 4) En fonction de la valeur du ratio, si l'arbre 2 est jugé meilleur ou non, la chaine prend l'arbre 2 comme valeur ou non. Cela est répété un grand nombre de fois, jusqu'à ce que la chaîne soit dans un état stationnaire, plus la chaîne passera de temps sur un arbre et plus sa probabilité postérieure sera élevée.

Théorie du coalescent

La méthode de coalescence est un processus généalogique permettant de retracer les relations de parenté entre des gènes en remontant le temps au cours des générations jusqu'à arriver à l'ancêtre commun le plus récent appelé MRCA. Aussi appelée coalescence de Kingman, elle a été développée dans les années 80 (Kingman 1982) et fonctionne de de façon à ce que chaque génération les lignées fusionnent jusqu'à n'en obtenir qu'une. Ce processus se fait sous des conditions de panmixie, apparentements aléatoires et évolution neutre des mutations et permet de calculer la vraisemblance des données sous un grand nombre de modèle de génétique des populations. Dans ces conditions pour une population haploïde de taille N, 2 lignées coalesceront en moyenne au bout de 2N générations. Après avoir généré l'arbre généalogique, les mutations supposées neutres sont déposées sur l'arbre permettant d'estimer la probabilité des données selon différents modèles. Le paramètre θ est une mesure de la diversité génétique de la population, dans la méthode de coalescence, $\theta = 4N\mu$ où μ est le taux de mutation par site et par génération. θ peut ensuite être estimé par méthode Bayésienne ou maximum likelihood sous les différents modèles de substitutions vus ci-haut. Dans une généalogie obtenue sous coalescence, chaque paire de lignées coalesce à un taux de 1/2N. Ainsi, si la taille de la population N varie au cours du temps, alors les lignées vont fusionner plus rapidement si N est petit et au contraire plus lentement si N est grand. De cette propriété on peut inférer, d'après l'allure de la généalogie, l'histoire démographique de la population. Ainsi différents modèles démographiques peuvent être utilisés afin d'inférer des changements de taille de populations passés. Le piecewise linear modèle considère par exemple que la fonction $\theta(t)$ est continue et « linéaire par morceaux », où $\theta(t)$ peut changer d'une fonction linéaire à une autre à plusieurs points dans le temps. Implémenté dans le programme Beast, celui-ci se nomme « Bayesian skyline plot » (Drummond et al. 2005), c'est un modèle largement utilisé pour les inferences démographiques, notamment dans cette thèse.

Bibliographie

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**: 1185–1192.

Hasegawa M, Kishino H, Yano T aki. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mamm Protein Metab* 21–123.

Kingman JFC. 1982. The coalescent. *Stoch Process their Appl* **13**: 235–248.

Ziheng Yang 2014. *Molecular Evolution :A Statistical Approach.*

Abstract

According to a 2015 WHO report, tuberculosis remains one of the top 10 causes of death worldwide. Despite considerable efforts by the United Nations to eradicate the disease by 2030, a global TB epidemic still persists. Its causative agent, the bacterium *Mycobacterium tuberculosis*, an obligate pathogen, has been plaguing humanity since it originated, and has coevolved with its main host, *Homo sapiens*, over thousands of years. Contemporary tuberculosis strains exhibit a structured phylogeographic pattern, carrying the genetic print of their geographic origin. The Koch bacillus infects and kills in large numbers, in poor and developing countries, where fragile health care systems, combined with high HIV prevalence, facilitate epidemic spread. In western countries, the major current threats are the multiplication and propagation of antibiotic resistant strains (MDR/XDR) coming predominantly from former Soviet republics.

In this thesis, I unravel the evolutionary history, propagation, and acquisition of drug resistance-conferring mutations in different settings, by implementing multiple genetic and genomic data sets. First, focusing on Central Asia, using whole genome sequencing and Bayesian statistics, I assess the effects of a treatment campaign on the development of MDR strains and highlight key mutations in successful strains. More importantly, the success of DOTs campaigns was compromised by the genetic make-up of these outbreak clades (pre-treatment low frequency resistance SNPs). Special attention was also given to a particular outbreak of MDR strains, i.e. the Russian W148 clone. I present its westward spatial and temporal propagation at a continental scale during the last century, and underline the key contribution of compensatory mutations in its epidemic success. However, tuberculosis does not only infect humans, but also has experienced successive mammalian host jumps. To decipher the adaptive constraints accompanying such secondary events, a systemic gene screen with selection signature-detecting algorithms was implemented to identify putative targets during diversifying selection.

Finally, novel mathematical tools and indices that reflect the epidemicity of a strain were developed, jumping from a population-driven approach to a strain specific one, with broader epidemiological applications. This allows us to correlate strain fitness with patient, lineage, and socio-economic information.

Résumé

D'après un rapport de l'OMS, la tuberculose reste en 2015 l'une des 10 premières causes de décès à l'échelle mondiale. De ce fait, en matière de santé, éradiquer la maladie à l'horizon 2030 est un des objectifs majeurs fixés par les Nations Unies. La bactérie responsable de cette infection, *Mycobacterium tuberculosis*, est un pathogène obligatoire dont l'origine et l'évolution sont intrinsèquement liées à celles de son hôte principal, *Homo sapiens*. En effet, les souches actuelles de tuberculose présentent, tout comme l'homme, une forte structure phylogénétique, trace de leur origine géographique. Les pays pauvres et en développement sont les plus touchés par l'épidémie globale, favorisée par des systèmes de santé défaillants et une haute prévalence du VIH. Les pays occidentaux ne sont pas épargnés, menacés par l'émergence de souches de plus en plus résistantes aux antibiotiques provenant en grande partie de l'ex URSS.

Au cours de cette thèse, j'analyse l'histoire évolutive, la propagation et l'acquisition de résistances aux antibiotiques de plusieurs épidémies de tuberculose en me basant sur des données génétiques et génomiques. Dans un premier temps je m'intéresse aux effets d'une campagne nationale de traitements en Asie Centrale sur le développement de souches multi-résistantes et met également en lumière le rôle clef de certaines mutations dans le succès des clones présentés. Ainsi cette campagne a été partiellement mise en échec par la présence de souches pré-résistantes, grâce à la survenue de mutations avant même la mise en place des traitements antibiotiques. Par la suite je me suis focalisé sur un clade particulier de souches multi-résistantes, le clone Russe W148. Je présente sa dispersion géographique et temporelle à travers l'Eurasie et démontre l'importance des mutations compensatoires dans son succès épidémique. De plus, la tuberculose ne touche pas seulement les hommes mais infecte également plusieurs autres mammifères. Afin d'appréhender les contraintes adaptatives accompagnants ces changements d'hôtes, j'ai effectué divers tests de sélection dans le but d'identifier les gènes impliqués.

Pour finir, nous avons développé un indice souche spécifique, permettant de mesurer le succès épidémique de celles-ci à un niveau individuel. Dans le cadre d'études épidémiologiques, cette mesure peut être croisée avec des informations sur le patient, la souche ou même socio-économiques.

Résumé

D'après un rapport de l'OMS, la tuberculose reste en 2015 l'une des 10 premières causes de décès à l'échelle mondiale. De ce fait, en matière de santé, éradiquer la maladie à l'horizon 2030 est un des objectifs majeurs fixés par les Nations Unies. La bactérie responsable de cette infection, Mycobacterium tuberculosis, est un pathogène obligatoire dont l'origine et l'évolution sont intrinsèquement liées à celles de son hôte principal, Homo sapiens. En effet, les souches actuelles de tuberculose présentent, tout comme l'homme, une forte structure phylogénétique, trace de leur origine géographique. Les pays pauvres et en développement sont les plus touchés par l'épidémie globale, favorisée par des systèmes de santé défaillants et une haute prévalence du VIH. Les pays occidentaux ne sont pas épargnés, menacés par l'émergence de souches de plus en plus résistantes aux antibiotiques provenant en grande partie de l'ex URSS. Au cours de cette thèse, j'analyse l'histoire évolutive, la propagation et l'acquisition de résistances aux antibiotiques de plusieurs épidémies de tuberculose en me basant sur des données génétiques et génomiques. Dans un premier temps je m'intéresse aux effets d'une campagne nationale de traitements en Asie Centrale sur le développement de souches multi-résistantes et met également en lumière le rôle clef de certaines mutations dans le succès des clones présentés. Ainsi cette campagne a été partiellement mise en échec par la présence de souches pré-résistantes, grâce à la survenue de mutations avant même la mise en place des traitements antibiotiques. Par la suite je me suis focalisé sur un clade particulier de souches multi-résistantes, le clone Russe W148. Je présente sa dispersion géographique et temporelle à travers l'Eurasie et démontre l'importance des mutations compensatoires dans son succès épidémique. De plus, la tuberculose ne touche pas seulement les hommes mais infecte également plusieurs autres mammifères. Afin d'appréhender les contraintes adaptatives accompagnants ces changements d'hôtes, j'ai effectué divers tests de sélection dans le but d'identifier les gènes impliqués. Pour finir, nous avons développé un indice souche spécifique, permettant de mesurer le succès épidémique de celles-ci à un niveau individuel. Dans le cadre d'études épidémiologiques, cette mesure peut être croisée avec des informations sur le patient, la souche ou même socio-économiques.

Mots Clés

Tuberculose, Génomique des populations, XDR/MDR, Sélection, Résistance aux antibiotiques, Phylogénies

Abstract

According to a 2015 WHO report, tuberculosis remains one of the top 10 causes of death worldwide. Despite considerable efforts by the United Nations to eradicate the disease by 2030, a global TB epidemic still persists. Its causative agent, the bacterium Mycobacterium tuberculosis, an obligate pathogen, has been plaguing humanity since it originated, and has coevolved with its main host, Homo sapiens, over thousands of years. Contemporary tuberculosis strains exhibit a structured phylogeographic pattern, carrying the genetic print of their geographic origin. The Koch bacillus infects and kills in large numbers, in poor and developing countries, where fragile health care systems, combined with high HIV prevalence, facilitate epidemic spread. In western countries, the major current threats are the multiplication and propagation of antibiotic resistant strains (MDR/XDR) coming predominantly from former Soviet republics. In this thesis, I unravel the evolutionary history, propagation, and acquisition of drug resistance-conferring mutations in different settings, by implementing multiple genetic and genomic data sets. First, focusing on Central Asia, using whole genome sequencing and Bayesian statistics, I assess the effects of a treatment campaign on the development of MDR strains and highlight key mutations in successful strains. More importantly, the success of DOTs campaigns was compromised by the genetic make-up of these outbreak clades (pre-treatment low frequency resistance SNPs). Special attention was also given to a particular outbreak of MDR strains, i.e. the Russian W148 clone. I present its westward spatial and temporal propagation at a continental scale during the last century, and underline the key contribution of compensatory mutations in its epidemic success. However, tuberculosis does not only infect humans, but also has experienced successive mammalian host jumps. To decipher the adaptive constraints accompanying such secondary events, a systemic gene screen with selection signature-detecting algorithms was implemented to identify putative targets during diversifying selection. Finally, novel mathematical tools and indices that reflect the epidemicity of a strain were developed, jumping from a population-driven approach to a strain specific one, with broader epidemiological applications. This allows us to correlate strain fitness with patient, lineage, and socio-economic information.

Keywords

Tuberculosis, Population genomics, XDR/MDR, Selection, Antibiotic resistance, Phylogeny